Phylogenetic and Developmental Studies into the Evolution of an Insect Novelty

Andrew David Economou

Research Department of Genetics, Evolution and Environment

UCL

Submitted for the Degree of Doctor of Philosophy

August 2008

Declaration

I, Andrew David Economou, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:

Abstract

The insects possess one of the most instantly recognisable bodyplans. This thesis addresses the evolution of one characteristic feature of the insects: the intercalary segment of the head. This small, appendageless segment is the homologue of the ancestral crustacean second antennal segment and its evolution underlies the loss of the second pair of antennae in the insect head.

There is little consensus between different methods of phylogenetic reconstruction as to which crustacean group the insects are most closely related to. This question is addressed by compiling a multigene dataset and running a number of Bayesian phylogenetic analyses to investigate the effects of analysing the data under different models of evolution. In addition, Bayes factor hypothesis tests addressing the position of the insects within the Pancrustacea are described.

The rest of the thesis addresses the developmental changes underlying the evolution of the intercalary segment. Almost everything that is known about the development of this segment in the insects comes from *Drosophila*. However, it is not clear exactly what constitutes the segment in the fly embryo. Specifically, it is unclear whether a pair of lobes behind the *Drosophila* stomodeum – the hypopharyngeal lobes – belong to the intercalary or mandibular segment. Using a detailed comparison of expression patterns between *Drosophila* and the red flour beetle *Tribolium*, the segmental affinity of these lobes is resolved.

Finally, a screen to identify potential candidate genes for patterning the intercalary segment is described. The screen makes use of the Berkley *Drosophila* Genome Project expression pattern database to identify genes expressed in the segment of the fly. Having identified orthologues of the genes in *Tribolium* using the genome sequence on BeetleBase, their expression patterns are examined in the beetle. Genes with conserved expression are deemed good candidates for a more widespread role in patterning the segment.

Acknowledgements

I would like to thank my supervisor Max Telford for all the help, guidance and opportunities he has given me over the past few years. I would also like to thank all the members of the Telford lab past and present; I am grateful to Rob Lanfear, Omar Rota-Stabelli, Daniel Papillon and Josh Coulcher for many lively and enthusiastic discussions, and especially to Sarah Bourlat for all the time and assistance given to me at the start of my studies and to Niko Prpic for his endless help and guidance with my embryological work.

Many thanks go to Ernst Wimmer and Gregor Bucher in Göttingen for many useful discussions regarding insect head development and for providing the beetles and flies, and to Chuck Cook, Jean Deutsch and Andrew Peel for providing various invaluable specimens and clones. Also to Linda Partridge for allowing me to use the fly facility and to all members of the Partridge lab for constant advice. Particular thanks go to Giovanna Vinti for the time and help given to me when I was starting my fly work. I would also like to thank all the members of the ZOONET labs for providing such a stimulating atmosphere to present my work and for many thought-provoking discussions.

Finally, I would like to thank all my friends and family for the support these past few years, and I would like to thank Chloe McCann, Josh Coulcher, Tom Osborne, my brother Charles and my mum for excellent proof reading.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Evo-devo and the evolution of morphology

1.1.1 Structure, function and morphology

In the closing lines of *The Origin of Species*, Charles Darwin described the evolution of "endless forms most beautiful and most wonderful" (Darwin, 1859). Understanding this diversity of organismal form has long provided fertile ground for biological enquiry. At the turn of the nineteenth century the great German polymath Johann Wolfgang von Goethe and the French naturalist Étienne Geoffroy Saint-Hilaire independently conceived of the notion of studying structural correspondences between the forms of different organisms. The significance of this "structuralist" view of morphology is seen most clearly when contrasted with the alternative "functionalist" view; these two perspectives are perhaps best exemplified by the contrasting views of Geoffroy and another great French naturalist George Cuvier, which lay behind one of the most famous and vigorous debates in biology.

Cuvier saw that animals shared distinct structural plans. Most notably he grouped the animals into his four *embranchements* (Vertebrata, Articulata, Molluska and Radiata) based on four distinct nervous systems. His structural groupings represented different functional needs; for example all vertebrates have similar structures because they carry out a similar set of functions. Importantly, for Cuvier the different morphologies represented by the *embranchements* were completely unrelated, so any comparison between them was essentially meaningless (Amundson, 2005, Hall, 1996). In summary,

an organism's structure was entirely the result of its function; the different structural plans represented groups of organisms carrying out similar functions.

Geoffroy also argued that the morphologies of all organisms conformed to structural plans. However, his notion of a structural plan differed from Cuvier's. Geoffroy argued that different organisms were composed of the same elements, and homologous elements could be found between the different organisms. For example he proposed that a mammalian shoulder girdle and a fish pectoral fin possessed homologous elements. Indeed, he believed that there was one archetype from which all animal morphologies could be derived (Amundson, 2005, Hall, 1996). In summary, an organism's morphology was a variant on a structural plan and function was secondary. This clearly opposed Cuvier's view that an animal's morphology was entirely dependent on its functional needs. A stormy series of eight debates before the Académie Royale des Sciences ensued between them, in which Geoffroy argued for homologies between Cuvier's distinct *embranchements*.

Cuvier's ideas were incompatible with any form of change between different morphologies; if environments changed, species would go extinct. For Geoffroy, if environments changed, the elements within a structural plan could adapt (Hall, 1996). After the publication of *The Origin of Species* in 1859, the idea of change between organisms with different morphologies became accepted; organisms were related through descent with modification. This provided a framework in which the structuralist perspective championed by Geoffroy could be understood. Homologous structures exist between different organisms as they have undergone different modifications during their separate descents from the common ancestor. Studying morphology in this manner became the dominant approach to studying evolution in the following decades (Amundson, 2005). This led to the establishment of questions about how different morphologies are related: which structures are homologous between different organisms and how did they differentiate?

1.1.2 Evo-devo and the evolution of bodyplans

This structuralist approach to evolution was common during the late nineteenth and early twentieth centuries. During the twentieth century, with the advent of the Modern Synthesis (the union of Darwinian natural selection with Mendelian genetics), evolutionary biology came to be dominated by population genetics (Gould, 2002). In the 1930s J. B. S. Haldane, R. A. Fisher and Sewell Wright formulated mathematical theories as to how genes would spread in populations and in the following decades several studies on natural populations were carried out to validate these theoretical predictions (Arthur, 2004). However, in the past few decades the rise of evolutionary developmental biology – or evo-devo as it is commonly known – has breathed new life into the structuralist approach to the study of morphological evolution (Hall, 2003).

Evo-devo is concerned with comparing development between different organisms (the process by which the morphology of an individual is built), to understand how changes in development lead to evolutionary changes in the phenotype. These studies have investigated a broad range of issues in morphological evolution such as the loss of eyes in cave dwelling forms of a single fish species (Yamamoto, *et al*., 2004) and the evolution of wing spots in different *Drosophila* species (Gompel, *et al*., 2005). Evo-devo studies have also compared development across much greater phylogenetic distances. Perhaps most notable have been the attempts to infer the form of the ancestral bilaterian and try to understand how it diversified into the range of morphologies seen across the Metazoa today (for example Hejnol and Martindale, 2008)

One of the most important concepts for these broad phylogenetic evo-devo studies has been that of the bodyplan. The essence of this concept is summed up nicely by Valentine and Hamiltion (1998) who describe a bodyplan as “the assemblage of morphological features that is found among members of a higher taxon”. Understanding how morphology has evolved at this broad phylogenetic level can be seen as trying to understand how different bodyplans have evolved. Bodyplans can differ greatly, however – compare for example the morphology of an arthropod with a

vertebrate – and so when addressing bodyplan evolution, it is first necessary to understand how the conserved bodyplans of the different taxa are structurally related.

Evo-devo studies have helped to resolve such questions about homology. For example, comparisons of gene expression domains have helped to resolve the homology of arthropod head segments (Damen, *et al.*, 1998, Telford and Thomas, 1998). Only once such homologies have been established can the developmental comparisons between the homologous structures give insight into the changes underlying the morphological differences. It is perhaps in these broader comparative studies that evo-devo bears most resemblance to the nineteenth century structuralist approach.

1.1.3 The role of phylogeny in evo-devo

Modern evo-devo studies are carried out in a robust phylogenetic framework (Hall, 2003). Only by mapping the different character states of a homologous structure onto a phylogeny can a morphological transition be defined (Telford and Budd, 2003). It is perhaps less obvious that an established phylogeny is also necessary to understand the developmental changes that occurred during a given morphological transition.

It is not enough to compare the development of one taxon exemplifying the ancestral state and one the derived state, especially when looking at a character conserved across a bodyplan. There are several examples to suggest that the development of a phenotypically conserved structure can vary between taxa. For example, in the fruit fly *Drosophila melanogaster*, the leg patterning gene *Distal-less* (*Dll*) is repressed by the two hox genes expressed in the abdomen: *Ultrabithorax* (*Ubx*) and *abdominal-A* (*abd-A*) (Vachon, *et al.*, 1992). It has therefore been suggested that the loss of legs in the insect abdomen is the result of repression from both these genes (Levine, 2002). However, in the red flour beetle *Tribolium castaneum* only *abd-A* represses appendage development; *Ubx* does not (Lewis, *et al.*, 2000). This shows that the use of a single taxon as an exemplar can be misleading. Jenner (2006) argues that even the common practice of choosing a supposedly underived basal taxon as an exemplar is "metaphysically" flawed.

Given this potential diversity in development even within a conserved morphological structure, the role of an established phylogeny is illustrated in figure 1.1. When there is a clade displaying a derived character state, nested within a larger group displaying the ancestral state, a phylogenetic framework is needed to infer the ancestral developmental pathways at both ends of the stem lineage (the lineage of extinct taxa leading to an extant "crown group") within which the morphological transition occurred. Only once these developmental pathways have been inferred can the developmental changes associated with the transition itself be inferred.

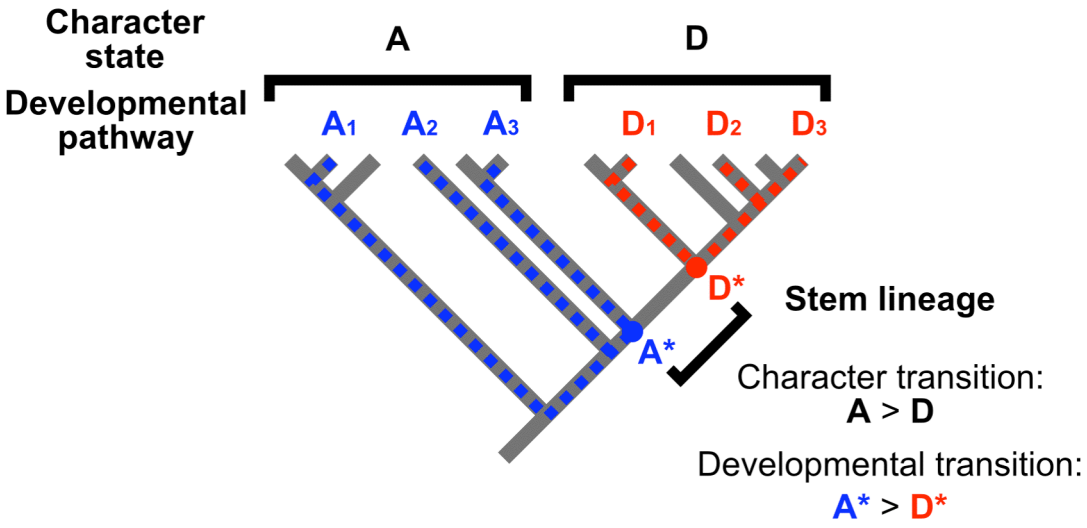


Figure 1.1. The importance of phylogeny in evo-devo. The schematic represents a clade showing a derived character state (D) nested within a larger group showing the ancestral state (A). The transition from the ancestral state to the derived state (A to D) occurs within the stem lineage. The developmental transition underlying this character transition can be inferred by comparing the developmental pathway for the ancestral character state at the base of the stem lineage (A*) with the developmental pathway for the derived state at the top of the stem lineage (D*). The stem lineage is composed of extinct organisms, so these developmental pathways must be inferred from the developmental pathways in extant organisms. The developmental pathways of various extant taxa are known: A_1, A_2 and A_3 for the ancestral character state and D_1, D_2 and D_3 for the derived character state. As the developmental pathways producing a conserved morphological character can vary it is necessary to compare across A_1, A_2 and A_3 to infer A*, and D_1, D_2 and D_3 to infer D*. However, as is indicated by the dashed lines, the developmental pathways in these extant taxa differ in how closely related they are to each other and in the case of A_1, A_2 and A_3, how closely related they are to A*. As it is most parsimonious to expect a feature of development shared by two organisms to be preset in their common ancestor, a knowledge of the phylogeny is needed to infer A* and D*.

1.2 The evolution of the insect bodyplan

1.2.1 Studying insect bodyplan evolution

Without doubt, one of the most instantly recognisable bodyplans is that of the insects. The insects are one of the most well known groups in terms of their ecological dominance, making up over half of all named species and occupying almost every conceivable terrestrial and freshwater habitat (Brusca and Brusca, 2003, Grimaldi and Engel, 2004). This diversity is found within a strongly conserved bodyplan (the main features of which are illustrated in figure 1.2). The insects have a head with a single pair of antennae and three pairs of mouthpart appendages (although these may be considerably modified for different modes of feeding), a thorax with three pairs of uniramous (unbranched) walking legs, and a legless abdomen (although some basal insects have various styli on at least some abdominal segments) (Richards and Davies, 1977). Insects also share a number of other features such as Malpighian tubules for osmoregulation and a tracheal system for breathing, and their embryos contain the amnion and serosal membranes.

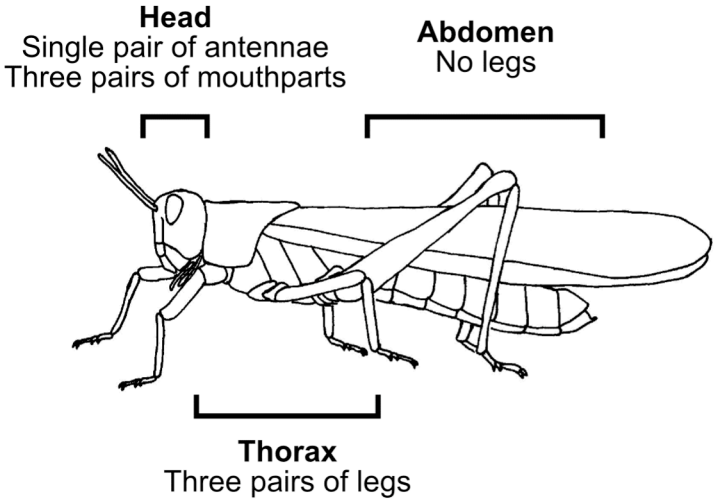


Figure 1.2. Illustration of the insect bodyplan. The major features of the insect bodyplan are clearly shown in an insect such as a locust. The body is divided into three parts: a head bearing a single pair of antennae and three pairs of mouthparts, a thorax bearing three pairs of uniramous (unbranched) legs and a legless abdomen (although some basal insects have various styli on at least some abdominal segments).

In terms of evo-devo, the insect bodyplan is of particular interest. As was illustrated above, when studying bodyplan evolution it is necessary to understand how the bodyplans of different organisms are structurally related through different lines of descent from a common ancestor, and the insect bodyplan is particularly suited to this type of study. Like all arthropods, insects are segmented organisms. The three major regions of the insect bodyplan (the head, thorax and abdomen) are groups of like segments that form functional units, known as tagmata (Brusca and Brusca, 2003). The thorax, for example, is a set of three segments each bearing a pair of appendages specialised as legs. This view of the bodyplan can be extended to all the other arthropod groups (Brusca and Brusca, 2003). The crustaceans (a diverse assemblage including familiar forms such crabs, water fleas and barnacles), the chelicerates (of which spiders and scorpions are the best known members) and the myriapods (millipedes, centipedes and some lesser known groups) all have bodyplans that can largely be defined by different patterns of tagmosis (see figure 1.3).

Clearly this view of arthropod bodyplans is an oversimplification. There are many important bodyplan features that cannot be accounted for by patterns of tagmosis, such as the Malpighian tubules or the tracheal system of the insects. However, viewing the evolution of the various arthropod groups in terms of their patterns of tagmosis sets up a clear framework to understand how some of the most important features of the various bodyplans evolved. How does segment number change, how are segments grouped into tagmata and within these tagmata how do segments evolve their particular specialisations? This framework makes studying the evolution of the insect bodyplan particularly appealing.

1.2.2 Insect developmental systems

There is perhaps a more critical feature that makes the insect bodyplan an attractive system to study. Any evo-devo study needs organisms that are amenable to developmental investigation. A number of different insects spanning the whole group have been used for developmental studies (see figure 1.4). Apart from the developmental model organism *Drosophila* (Diptera), a number of sophisticated

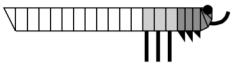
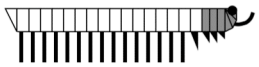
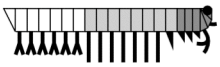
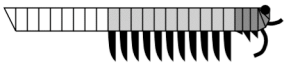
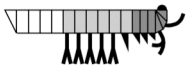
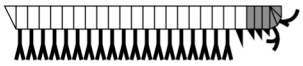
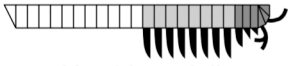
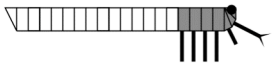
	Tagma (and segment number)	Appendage type
Insect  *Lepisma* (Silverfish)	Head (5) Thorax (3) Abdomen (up to 11)	All uniramous Abdominal appendages only present in basal groups
Myriapod  *Scutigera* (House centipede)	Head (5) Trunk (variable)	All uniramous Present on all segments
Crustacean		
Malacostracan  *Porcellio* (Woodlouse)	Head (5) Thorax (8) Abdomen (6)	Uniramous or biramous Present on all segments
Branchiopod  *Artemia* (Brine shrimp)	Head (5) Thorax (usually 11) Abdomen (variable)	Uniramous or biramous in head Usually phyllopodous in tunk Presence of abdominal appendages varies within group
"Maxillopod"  *Mesocyclops* (Copepod)	Head (5) Thorax (6) Abdomen (4)	Uniramous or biramous Abdominal appendages absent
Remipede  *Speleonectes*	Head (5) Trunk (variable)	Biramous Present on all segments
Cephalocarid  *Hutchinsoniella*	Head (5) Thorax (8) Abdomen (11)	Uniramous or biramous in head Phyllopodous in thorax Absent in abdomen
Chelicerate  *Buthus* (Scorpion)	Prosoma (6) Opisthosoma (up to 12)	Uniramous Presence of opisthosomal appendages varies within group

Figure 1.3. The different arthropod bodyplans. *(Previous page).* The main features of the bodyplans characterising the different arthropod groups are summarised. For each group the pattern of tagmosis is given, including the number of segments making up each tagma and the appendage types present (uniramous, biramous or phyllopodous) (based on Brusca and Brusca, 2003). For each group the bodyplan is represented with a schematic of a member of the group, illustrating how the types of appendages on the different segments vary along the body. The different tagmata are shown in different shades of grey. The bodyplans for the entognathous hexapods (collembolans, proturans and diplurans) are not shown as they are largely the same as the insect bodyplan. The myriapod bodyplan is represented by the chilopods which do not show the different numbers of tergites to sternites seen in the other myriapod groups (diplopods, symphylans and pauropods). There is no single crustacean bodyplan. Therefore, the bodyplans of the major crustacean subgroups are shown. For the malacostracans, the eumalacostracan bodyplan is shown; the bodyplan of the phyllocarids (basal malacostracanas) is largely the same although there are minor differences. The maxillopod bodyplan is largely shared by a number of crustacean groups, most notably the copepods, the cirripedes and arguably the ostracods.

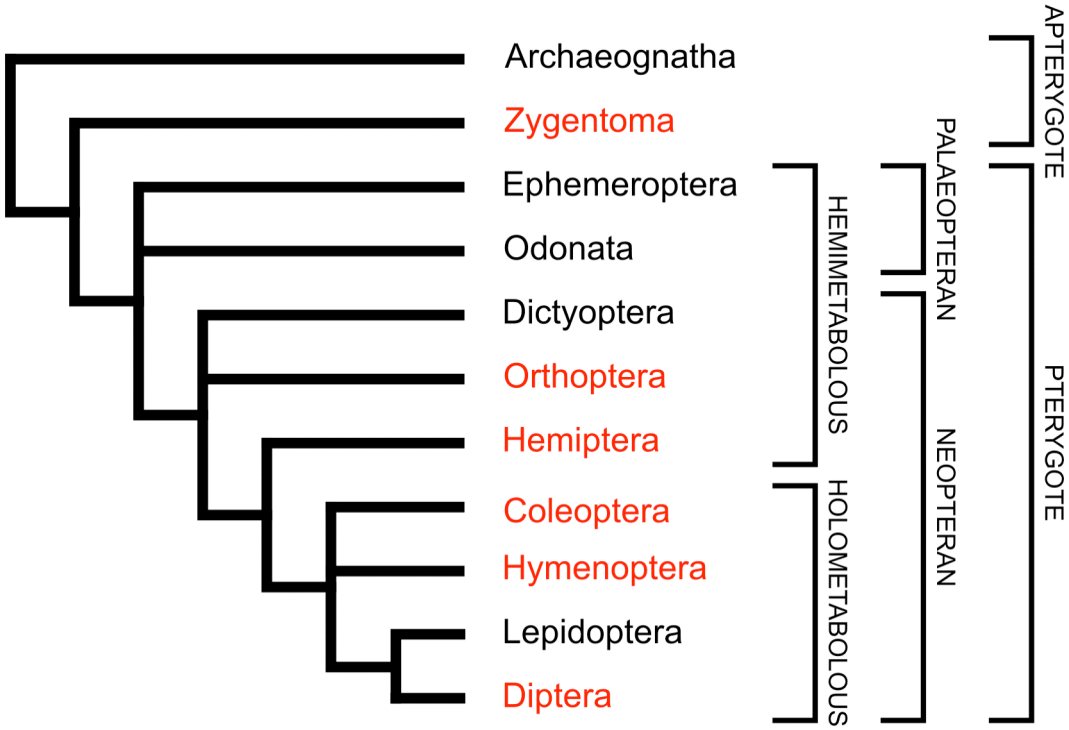


Figure 1.4. Insect phylogeny and the distribution of developmental systems. The phylogenetic relationships of some of the best known insect orders are shown. Orders containing insects that have been used for developmental study are marked in red. These orders represent a range of states for some of the important characters of insect morphology and development which vary across the group (shown on the right). These are the presence of wings (pterygote as opposed to apterygote), the ability to fold wings over the abdomen (neopteran as opposed to palaeopteran) and complete metamorphosis (holometabolous as opposed to hemimetabolous where nymphal stages resemble adults but are sexually immature and have wing buds). The phylogenetic relationships are largely based on Kristensen (1981) and Grimaldi and Engel (2004).

techniques such as transgenics are being developed for *Tribolium* (Coleoptera) (Klingler, 2004). The parasitoid wasp *Nasonia vitripennis* (Hymenoptera) and within the hemimetablolous insects the milkweed bug *Oncopeltus fasciatus* (Hemiptera) and the cricket *Gryllus bimaculatus* (Orthoptera) are amenable to simple functional studies using RNAi (for example Hughes and Kaufman, 2000, Lynch, *et al*., 2006, Miyawaki, *et al*., 2004). Even the basal apterygote insect *Thermobia domestica* (Zygentoma) has been used for studying expression patterns (for example Peterson, *et al*., 1999).

There are several other arthropods for which developmental techniques are also being developed. The crustaceans *Parhyale hawaiensis*, *Artemia franciscana* and *Daphnia pulex*, the myriapods *Strigamia maritima* and *Glomeris marginata* and the chelicerate *Cupiennius salei* have all been used for comparative developmental studies (for example Chipman, *et al*., 2004, Copf, *et al*., 2003, Papillon and Telford, 2007, Pavlopoulos and Averof, 2005, Prpic and Tautz, 2003, Stollewerk, *et al*., 2003). Therefore, not only is it eminently feasible to infer the ancestral mode of development for an insect character, it is also possible to make inferences for the mode of development of the homologous character in other arthropods.

It is important to point out that there are a few small groups of arthropods that I have not yet introduced, known as the entognathous hexapods, which share many of the features of the insect bodyplan (Richards and Davies, 1977). These taxa (the Collembola, Protura and Diplura) have essentially the same pattern of tagmosis as the insects and they are generally seen as the sister-taxa to the insects; together with the insects they form the Hexapoda (Luan, *et al*., 2005). Questions regarding the evolution of many features of the insect bodyplan can be extended into a larger hexapod bodyplan. I will generally not discuss this larger hexapod bodyplan. So far the entognathous hexapods have not proved amenable to developmental study so it is not practical to discuss inferring developmental states for characters shared across the hexapods. Whilst a large number of developmental systems makes the insects a good system for evo-devo studies, this cannot be said of the larger hexapod grouping.

1.3 Arthropod phylogeny and the insect bodyplan

1.3.1 Arthropod phylogeny and the position of the insects

It was shown above (in section 1.1.3) that any evo-devo study needs to be viewed in a phylogenetic context. Arthropod phylogeny has long proved an area of intense debate and much controversy has existed over the interrelationships of the four major arthropod classes: the hexapods (including the insects), the myriapods, the crustaceans and the chelicerates (for a review see Regier and Shultz, 1997). One long-standing area of agreement, however, was the grouping of the hexapods with the myriapods in a group named the Atelocerata (also Tracheata or Antennata). These two groups share a number of characters (Dohle, 1998, Kraus, 1998, Regier and Shultz, 1997): both have a head with a single pair of antennae, both lack multiramous (branched) appendages and both were argued to have 'telognathic' mandibles, where the mandibles bite at the tip. This is in contrast to crustaceans, which have two pairs of antennae and were described as having gnathobasic mandibles, where the base of the appendage handles the food (the distal portion of the appendage being reduced to a palp) and to the chelicerates, which do not possess antennae or mandibles. Both crustaceans and chelicerates also contain members with multiramous appendages. Additionally, the insects and myriapods share the tracheal system for breathing and osmoregulate using Malpighian tubules.

In the last decade and a half, this traditional view of a close relationship between the insects and myriapods has been challenged. A number of molecular phylogenetic analyses, based on a range of genes, have addressed the relationships of the arthropod taxa. These studies repeatedly uncovered evidence to support a close relationship between the insects and the crustaceans to the exclusion of the myriapods (Cook, *et al.*, 2001, Cook, *et al.*, 2005, Friedrich and Tautz, 1995, Giribet, *et al.*, 2001, Giribet, *et al.*, 2005, Hwang, *et al.*, 2001, Lavrov, *et al.*, 2004, Mallatt and Giribet, 2006, Mallatt, *et al.*, 2004, Nardi, *et al.*, 2003, Negrisolo, *et al.*, 2004, Pisani, *et al.*, 2004, Regier and Shultz, 1997, Regier and Shultz, 2001, Regier, *et al.*, 2005, Shultz and Regier, 2000, Spears and Abele, 1998, Turbeville, *et al.*, 1991). This grouping has also been supported by mitochondrial gene order (Boore, *et al.*, 1998).

In addition to the molecular evidence, this so called Pancrustacea (or Tetraconata) hypothesis has gained additional support from recent work looking at the nervous system. Ommatidial structure, the presence of neuronal stem cells, brain structure and patterns of serotonin-immunoreactive neurons have all been argued to support a grouping of insects and crustaceans (Harzsch, 2004, Harzsch, *et al.*, 2005). While this has strengthened the Pancrustacea hypothesis, several features supporting the Atelocerata have been refuted or questioned. Most notably work looking at the expression of the gene *Dll*, a marker for the distal parts of appendages, has demonstrated that insects and myriapods, like crustaceans, have gnathobasic mandibles, not telognathic mandibles as previously argued (Popadic, *et al.*, 1998, Popadic, *et al.*, 1996). It is also likely that several of the other features supporting the Atelocerata are the result of convergent evolution to a terrestrial mode of life, as basis for homology has been questioned (Dohle, 1998, Kraus, 1998).

1.3.2 Crustacean diversity and the insect bodyplan

The grouping of the insects with the crustaceans has major implications for the evolution of the insect bodyplan. Under the Atelocerata hypothesis this bodyplan would have been derived from a larger group bearing the many features shared between the insects and myriapods: a head with a single pair of antennae, uniramous legs, trachaea and Malpighian tubules. Understanding the evolution of the insects would have centred on how the segment number stabilised and how the distinctive pattern of tagmosis seen in the insects was derived most probably from a more homonomous bodyplan as seen in the myriapods.

Under the Pancrustacea hypothesis, the morphological transitions involved in the evolution of the insect bodyplan are much less clear. Firstly, the crustaceans are made up of a number of different subgroups. The most speciose of these are the Malacostraca (including a range of well known forms such as crabs, lobsters, woodlice and mantis shrimps) and the Branchiopoda (which include brine shrimps, water fleas and tadpole shrimps). The crustaceans also include a number of taxa that were previously grouped together as the “Maxillopoda” – now believed to be a polyphyletic group (Mallatt and

Giribet, 2006, Regier, *et al*., 2005, Wills, 1998) – the most important of these being the Cirripedia (the barnacles) and the Copepoda (a large marine radiation including many planktonic forms), as well as the Ostracoda (seed shrimps) which were also sometimes placed in the "Maxillopoda", and the enigmatic Remipedia and Cephalocarida. These groups have very different bodyplans (see figure 1.3), not just in comparison to the insects, but also to each other. Most notably, the patterns of tagmosis and the structure of the appendages differ greatly between the various groups.

It is not immediately obvious how the insects relate to this assemblage. There are no overwhelming crustacean synapomorphies which would exclude the insects from falling within the group (although some characters that support a monophyletic Crustacea are given inEdgecombe, 2004). Moreover, such is the diversity of crustacean morphology, that there has been little consensus between the many attempts to reconstruct crustacean phylogeny based on morphology alone (Wills, 1998). Whilst there may be overwhelming molecular and neurobiological evidence in support of an insect-crustacean clade, these methods have been unable to resolve precisely how the insects relate to the crustaceans. In the absence of an established phylogeny it is difficult to make any hypotheses for the character transitions involved in the evolution of insect tagmosis or appendage type as the immediate outgroup is not known.

There are other uncertainties associated with a crustacean origin for the insects. For a number of features the insects clearly show a derived state, but the crustacean homologue is unclear, such as the tracheal system and Malpighian tubules – although there have been some recent advances in this area (Franch-Marro, *et al*., 2006). The crustacean origin for the insects has, therefore, made several of the transitions involved in the origin of the insect bodyplan difficult to define. There is one insect feature where the transition from ancestral crustacean state to a derived state in the insects is clear, however, namely a segment in the insect head known as the intercalary segment.

1.4 The intercalary segment

1.4.1 The importance of the insect intercalary segment

In spite of the diversity of crustacean bodyplans, one feature that is conserved across the different crustacean groups is the presence of two pairs of antennae (Brusca and Brusca, 2003). This contrasts with the single pair seen in the insects. Comparisons of the expression of the segmental marker gene *engrailed* (*en*) between insects and crustaceans have shown that this is underpinned by a very simple difference (see figure 1.5). The crustacean head consists of a pregnathal head with two antennal segments (the segmental composition of the more anterior portions of the head is still debated) followed by the three mouthpart segments of the gnathal head (the mandibular segment and two pairs of maxillary segments) (Scholtz, 1995). This head structure is conserved in insects, except that the homologue of the second antennal segment is a small, appendageless segment called the intercalary segment (Scholtz, 1998).

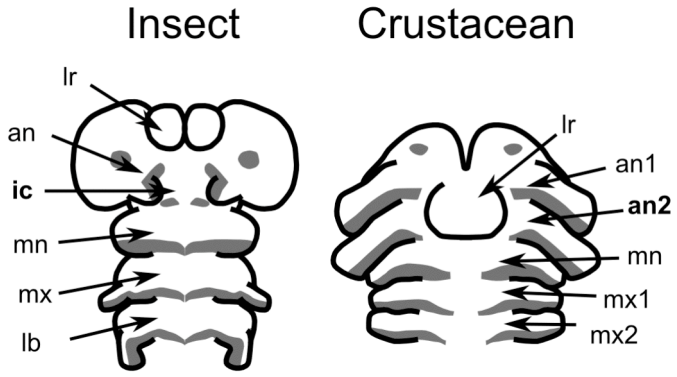


Figure 1.5. Comparison of the segmental compositions of insect and crustacean heads. Schematics depict *en* expression (grey) in the embryonic heads of an insect (based on *Tribolium castaneum*) and a crustacean (based on Parhyale hawaiensis, see Browne*, et al*., 2005) marking out the different segments. The segmental compositions of the heads are essentially the same: both have three pairs of mouthpart segments – a mandibular segment and two maxillary segments (the insect second maxillary segment is called the labial segment) – and both have an anterior antennal segment. The main difference is that where the crustaceans have a large appendage bearing second antennal segment, the insects have the small appendageless intercalary segment; these segments are marked in bold. an, antennal segment; ic, intercalary segment; lb, labial segment; lr, labrum; mn, mandibular segment; mx maxillary segment.

Despite the absence of a resolved pancrustacean phylogeny, the grouping of the insects with the crustaceans means that the ancestral state to the insect intercalary segment is the crustacean second antennal segment. This transition underlies the loss of the crustacean second antennae and so is one of the most characteristic transitions in the evolution of the insect bodyplan. Even if the insects are the sister-group of a monophyletic Crustacea rather than falling within a paraphyletic crustacean group, the intercalary segment must still have evolved from an appendage bearing segment and the crustaceans are the closest outgroup for comparison.

As was illustrated above, however, to describe fully the developmental changes behind the transition, it is still necessary to have a resolved pancrustacean phylogeny and to know the sister-group to the insects. Besides the importance of knowing how the insects relate to the different crustaceans for understanding the various other character transitions involved in the evolution of insect bodyplan, this is also necessary to understand the developmental changes behind the evolution of the intercalary segment.

1.4.2 Features of the intercalary segment

In order to describe the developmental changes underlying the transition from the second antennal segment to the intercalary segment, it is first necessary to clarify the precise morphological transformations that have occurred. A typical arthropod segment bears a number of features: there are paired appendages, mesodermal coelomic cavities (also known as somites), and neuromeres (Matsuda, 1965). The crustacean second antennal segment largely conforms to this canonical segmental structure (Anderson, 1973). For the insect intercalary segment, paired neuromeres are easily identifiable giving rise to the tritocerebrum of the insect brain (Harzsch, 2004). However, this is a plesiomorphic character seen across all the arthropods and is not related to the evolution of the intercalary segment. For the other features of a segment, the intercalary segment shows a clear derived morphology. I will now document these derived features of intercalary segment morphology.

Loss of appendages

As detailed earlier, probably the most striking feature of the intercalary segment is the lack of the pair of appendages seen in the ancestral crustacean second antennal segment. Whilst appendages are not present on the intercalary segment in the adult heads of any insect, a number of paired bulges seen on this segment in various insects have been described as transient appendages that are resorbed later in development (Roonwal, 1937, Tamarelle, 1984). This is most obvious in the immediate sister taxa to the insects, the entognathous hexapods (Ikeda and Machida, 1998, Tamarelle, 1984, Uemiya and Ando, 1987).

It has also been argued by some authors that the labrum represents the appendages of the intercalary segment (Haas, *et al*., 2001). This has been supported by various sources such as its innervation from the tritocerebrum – the neuromere belonging to the intercalary segment (Boyan, *et al*., 2002). However, it seems improbable that the labrum represents the appendages of the intercalary segment, as crustaceans possess a pair of antennae on their second antennal segment – the homologue of the intercalary segment – as well as possessing a labrum. In response to this criticism, it has been claimed that the labrum represents the endites of an intercalary appendage (or the second antenna) (Haas, *et al*., 2001). Whilst this appears unlikely, if it were true, the evolution of the intercalary segment would still involve the large-scale reduction of the appendage belonging to the segment, but the endites of the appendage would not have been lost in the insects.

Derived coelomic sacs

One of the most careful descriptions of mesoderm development in the classical literature is by Ullmann (1964). She describes the mesoderm of the intercalary segment of the beetle *Tenebrio molitor* as having a different histology to that of other segments, and the timing of the formation of the intercalary coelomic sacs differs from other segments. Ullmann (1964) also describes these intercalary sacs as giving rise to a transient embryonic structure known as the suboesophagal body, although other authors attribute this structure to the mandibular segment (Roonwal, 1937). It is not clear what

the crustacean homologue of the suboesophagal body is. De Velasco *et al*. (2006) also argue that a major embryonic derivative of the intercalary segment not recognised in the classical histological studies are hemocytes. The derivatives of the intercalary segment mesoderm are very different to those of a canonical segment.

A vestigial segment

In many ways, the intercalary segment appears to be a vestigial segment. It appears so reduced that until recently even its existence had been questioned (Singh, 1981). The expression of the segment polarity gene *en* in *Drosophila* and a range of other insects unequivocally demonstrated the existence of the segment (Diederich, *et al.*, 1991, Rogers and Kaufman, 1996, Schmidt-Ott and Technau, 1992). However, it is notable that here the *en* stripes are highly reduced in size and their onset delayed relative to those of other segments. Along with the loss of appendages, these observations fit in with the idea that the segment is largely vestigial when compared to its crustacean homologue.

The intercalary segment and adult head

So far, these descriptions of the insect intercalary segment have largely been restricted to embryological features. As Matsuda (1965) points out, in postembryonic stages the insect head is composite and compact. External structures, musculature and innervation show high degrees of fusion or reduction making it very hard to establish what structures belong to which of the different segments. Whilst there have been various theories, such as the insect hypophrarynx deriving from the intercalary segment, specifically from a pair of lobes called the hypopharygeal lobes (or *hypopharynxhöcker*) (Matsuda, 1965, Roonwal, 1937), these theories have often been questioned (for example Wolff and Scholtz, 2006). Therefore, whilst the differentiation of the intercalary segment into adult structures is clearly of great interest, it will not be discussed further.

1.4.3 Development of the intercalary segment

Very little is currently known about the genetics underlying the development of the intercalary segment. In *Drosophila*, the overlapping expression domains of the three head "gap-like" genes *orthodenticle* (*otd*), *empty spiracles* (*ems*) and *buttonhead* (*btd*) have been implicated in the establishment of the segments of the procephalon, including the intercalary segment (Cohen and Jürgens, 1990); this model of head segmentation is illustrated in figure 1.6. Aspects of this model have been questioned in *Drosophila* (Gallitano-Mendel and Finkelstein, 1998, Wimmer, *et al*., 1997), but recent work has clearly shown that these large overlapping domains do not define the head segments in *Tribolium* (Schinko, *et al*., 2008). Whilst the late expression and function of *Tc-otd1* resemble the gap-like role of *Drosophila otd*, *Tc-ems* expression and function does not span the range of segments seen in its *Drosophila* orthologue (being restricted to the posterior ocular region and anterior antennal segment), and *Tc-btd* is not required for head cuticle formation. This suggests that the role for the head gap-like genes in establishing the intercalary segment may not be conserved in other insects.

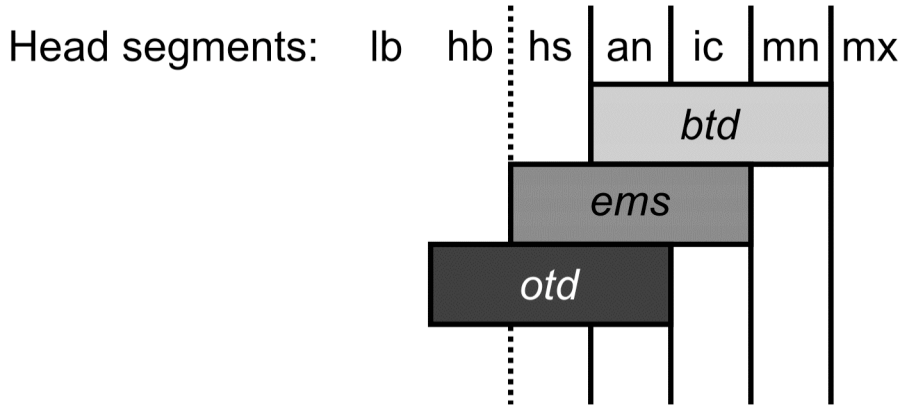


Figure 1.6. The role of the head gap-like genes in the establishment of head segments of *Drosophila*. The schematic represents the overlapping sets of segmental defects in cephalic "gap-like" mutants. *btd* mutants lose antennal, intercalary and mandibular segment structures and segment polarity gene expression. *ems* mutants lose antennal and intercalary segment structures and segment polarity gene expression, as well as some preantennal structures and the preantennal *en* head spot. *otd* mutants lose antennal segment structures and segment polarity gene expression, as well as some preantennal structures, the *en* head spot and the *wg* head blob. Therefore, it has been argued that the overlapping domains of the gene expression, shown in the schematic, are required for the establishment of the head segments: *btd* alone for the mandibular segment, *btd* and *ems* for the intercalary segment, *btd*, *ems* and *otd* for the antennal segment and *ems* and *otd* for the anterior head. Based on figure 3 from Cohen and Jürgens (1990). an, antennal; ic, intercalary; hb, head blob; hs, head spot; lb, labrum; mn, mandibular; mx maxillary.

A few more genes have been implicated with a role in patterning the segment in *Drosophila*. *knot* (*kn*) has been implicated in establishing the posterior boundary of the intercalary segment in the fly, seemingly working downstream of the head gap-like genes (Crozatier and Vincent, 1999). Two further genes, *cap'n'collar* (*cnc*) and *crocodile* (*croc*) have been implicated along with *kn* in the differentiation of the ventral intercalary segment of *Drosophila* (Häcker, *et al*., 1995, Rogers and Kaufman, 1997, Veraksa, *et al*., 2000). However, the roles of these genes in patterning the intercalary segment have been questioned (Mohler, *et al*., 1995, Seecoomar, *et al*., 2000). It is unclear whether or not the part of the embryo in which these genes are expressed and which they pattern, is in fact part of the intercalary segment.

The hox genes have been implicated in controlling the identity of several segments in the arthropods (Hughes and Kaufman, 2002b) and the anterior-most two hox genes *labial* (*lab*) and *proboscipedia* (*pb*) are expressed in the intercalary segment. However, neither gene has an obvious role in patterning the segment. *lab* is expressed in the segment throughout the insects (although this in contentious in *Drosophila*) (Angelini, *et al*., 2005, Diederich, *et al*., 1989, Nie, *et al*., 2001, Peterson, *et al*., 1999), but, where investigated, the function of this gene is unclear. In both *Drosophila* and *Oncopeltus* there is no obvious phenotype in *lab* mutants or RNAi knock downs relating to the intercalary segment (Angelini, *et al*., 2005, Merrill, *et al*., 1989). *pb* is also expressed in the intercalary segment of several insects (although not in *Drosophila*): transcripts accumulate in the intercalary mesoderm late in development, but no intercalary phenotype has been reported from any functional work on *pb* (Hughes and Kaufman, 2000, Rogers, *et al*., 2002, Shippy, *et al*., 2000).

Interestingly, co-expression of the head gap-like genes *ems* and *btd* has been implicated in giving the intercalary segment its identity in *Drosophila* (Schöck, *et al*., 2000). This suggests that the identity of the intercalary segment may not be conferred by the hox genes, but rather by the combinatorial action of the head gap-like genes. However, as has already been seen, the functions of the *Drosophila* head gap-like genes are unlikely to be conserved in other insects.

In summary, little is known about intercalary segment development. For the genes with conserved expression patterns in the intercalary segments of a range of insects, their roles in patterning the segment are unclear (*lab* and *pb*). Moreover, whilst a number of other genes have been implicated in patterning the segment in *Drosophila* and their functions are better understood, they either seem to be not conserved in other insects (*ems* and *btd*), or it is unclear whether they are actually involved in patterning the *Drosophila* intercalary segment (*kn*, *cnc* and *croc*).

1.5 Aims and objectives

In this thesis I investigate the evolution of the insect intercalary segment. As I have demonstrated, there is not a well-established phylogenetic framework in which to make developmental comparisons and little is currently known about the development of this segment in the insects. My investigations, therefore, relate to two main objectives: establishing a resolved phylogeny for the Pancrustacea and advancing the understanding of intercalary segment development.

In chapter 3 I address the first of these objectives. I apply the increasingly common approach of using a multigene dataset to analyse the phylogeny of the Pancrustacea. I compile and analyse a large dataset consisting of genes previously used to in arthropod phylogenetics as well as some newly sequenced genes, addressing various uncertainties in how to analyse such data. Using the most appropriate method of analysis I run a series of hypothesis tests investigating the support for different positions of the insects.

In chapters 4 and 5 I address the second objective. Firstly, it is unclear what constitutes the intercalary segment in the model organism *Drosophila melanogaster* from where most of what is known about the development of the segment comes. In chapter 4 I present a comparative study of gene expression patterns between *Drosophila* and the red flour beetle *Tribolium castaneum* aimed at resolving this issue.

In chapter 5 I attempt to expand the number of candidate genes for patterning the intercalary segment. I present a screen using *Drosophila* and *Tribolium* aimed at finding novel genes with conserved expression in the intercalary segment, as these would appear likely candidates for a role in the development of the segment.

Chapter 2:
Materials and Methods

I will now present the general methods used for this thesis, namely molecular cloning and sequencing (section 2.1), phylogenetic techniques for the analysis of pancrustacean phylogeny (section 2.2) and embryological techniques for studying the red flour beetle *Tribolium castaneum* and the fruit fly *Drosophila melanogaster* (section 2.3). Specific modifications to these general methods are documented within the individual chapters.

2.1 Molecular cloning and sequencing

In order to sequence genes for molecular phylogenetic analysis (chapter 3) or to synthesise probes for *in situ* hybridisation (chapters 4 and 5) the genes of interest had to be cloned into plasmid vectors. The fragments of interest were first amplified by the polymerase chain reaction (PCR). Reagents that did not belong to a kit were made according to Sambrook and Russell (2001), unless stated otherwise.

2.1.1 Polymerase chain reaction

A standard set of PCR conditions were used to amplify fragments for cloning. Reactions were carried out using the Roche Taq DNA Polymerase set (DNA Taq polymerase and 10x buffer; Cat. No. 1 596 594), the AB gene dNTP set (high

concentration; Cat. No. AB-0315) and with primers ordered from Thermo Electron or MWG. dNTP and primer stocks were diluted to the given concentration with Milli-Q water. Reactions were carried out in a total volume of 30 μl with the following volumes of reagents:

3.0 μl 10x buffer
22.8 μl Milli-Q water
1.0 μl dNTP (5mM)
1.0 μl Forward primer (10nM)
1.0 μl Reverse primer (10nM)
1.0 μl DNA
0.2 μl Taq DNA polymerase

Concentrations of the DNA samples varied. In several cases where PCRs only yielded a small amount of product, increasing the amount of DNA added to 2.0 μl often improved the yield.

PCR reactions were carried out in a Bio-Rad iCycler, Applied Biosystems GeneAmp PCR system 2700 or a G-Storm Thermal Cycler. Reactions were carried out in batches containing different sets of primer pairs. Different primer pairs with different melting temperatures were often run simultaneously. To ensure annealing in all reactions the annealing temperature was set to 2ºC below the lowest melting temperature (T_m) for any of the primers in one batch. Melting temperatures were estimated by the simple empirical rule known as "the Wallace rule" (Sambrook and Russell, 2001):

$$T_m \text{ (in ºC)} = 2(A+T) + 4(G+C)$$

where A, T, C and G are the number of each base in the oligonucleotide. In some cases, certain primer pairs only gave a small amount of product. Often the melting temperatures of these primer pairs were much higher than the annealing temperature used. Repeating the reaction with an annealing temperature closer to the melting temperature regularly improved the efficiency of the reaction. Extension times were matched to the estimated fragment length, assuming a transcription rate of 1000 bp/min.

Unless stated otherwise, a basic PCR cycle was used consisting of 1 cycle extended DNA denaturation of 2 min at 94ºC, followed by 35 cycles of 30 sec denaturation at 94ºC, 30 sec annealing at a temperature as calculated above and extension for the appropriate length of time at 72ºC, followed by a final extension step of 10 min at 72ºC.

2.1.2 Reverse Transcriptase PCR

RNA samples had to be reverse transcribed into cDNA before fragments could be amplified. This was done by Reverse Transcriptase PCR (RT-PCR), using a protocol kindly provided by Dr Sarah Bourlat. First the RNA was denatured. 0.5 µl RNA was mixed with 2.5 µl Milli-Q water and heated to 65ºC for 10 min in a thermocycler. This was followed by first strand synthesis using the Roche Expand Reverse Transcriptase set (Expand Reverse Transcriptase, 5x concentration Expand Reverse Transcriptase buffer and dithiothreitol (DTT)), Roche Hexanucleotide mix (Cat. No. 11785826001), dNTP (AB gene dNTP set (high concentration; as above) and Roche Protector RNase Inhibitor (Cat. No. 3335399). The hexanucleotide mix was diluted in Milli-Q water ten-fold before use. First strand synthesis was carried out in a total volume of 10 µl with reagents added to the denatured RNA in the following volumes:

2 µl 5x concentration Expand Reverse Transcriptase buffer
2.25 µl Milli-Q
0.5 µl DTT
0.5 µl Hexanucleotide mix
1 µl dNTP
0.25 µl RNase Inhibitor
0.5 µl Expand Reverse Transcriptase

First strand synthesis was performed in a thermocycler with annealing at 25ºC for 10 min, reverse transcription at 42ºC for 60 min and inactivation at 95ºC for 10 min.

The total 10 μl cDNA product was used for PCR, with reagents and cycling conditions as described in section 2.1.1. Reactions were carried out in a total volume of 25 μl and reagents were added in the following volumes:

10 μl	First strand synthesis product
2.5 μl	10x buffer
9.4 μl	Milli-Q water
1.0 μl	dNTP (5mM)
1.0 μl	Forward primer (10nM)
1.0 μl	Reverse primer (10nM)
0.1 μl	Taq DNA polymerase

For certain applications (see section 2.3.4), the entire cDNA was used not used in the PCR, but was stored as a stock. In these cases, the RT-PCR reaction was carried out in a total volume of 20 μl, with the volumes of reagents in the denaturation and first strand synthesis steps doubled. cDNA was stored at –20ºC.

2.1.3 PCR product isolation and purification

It was possible that the PCR had amplified fragments other than the one desired, especially if using degenerate primers (as in section 2.2.3). Therefore, to confirm that the PCR had amplified the desired fragment and to isolate that fragment, the PCR products were separated by agarose gel electrophoresis (see section 2.1.4). DNA was visualised on a UV light box and bands of the expected size were excised with a scalpel.

The excised DNA was purified from the gel using the QIAGEN QIAquick Gel Extraction Kit (Cat. No. 28706) or the QIAGEN MinElute Gel Extraction Kit (Cat. No. 28606). For both kits, DNA is adsorbed to a silica membrane in the presence of high concentrations of salt, whilst impurities and contaminants pass through the membrane. DNA is then eluted into an elution buffer. The QIAquick kit can produce 30 μl elutant, whilst the MinElute kit produces a more concentrated DNA extract in a volume of 10 μl. The more concentrated extract was used in cases when there were problems in

cloning that could have been due to DNA concentration. 2 μl of the purified PCR product were run on an agarose gel to confirm that the purification was successful.

2.1.4 Agarose gel electophoresis

Agarose gel electrophoresis was used to separate DNA fragments of different sizes. As the fragments of interest tended to be 0.5-1.5 kb, gels were made with 1% agarose in 1x TBE or 1x TAE. Ethidium bromide was added to the gel (approximately 1 μl (at 10 mg/ml) per 200 ml), allowing visualisation of DNA under UV light. A 1 kb ladder (Invitrogen 1 kb DNA Ladder; Cat. No. 15615-024) was also run with the samples, allowing the size of the DNA fragments to be judged.

2.1.5 Cloning

The purified PCR fragments were cloned into the TOPO TA cloning pCR II-TOPO vector (Cat. No. K4600-40) or Promega pGEM-T Easy vector (Cat. No. A1360). Both kits are suitable for cloning PCR products amplified with a Taq DNA polymerase. The TOPO TA cloning system uses a topoisomerase to insert the product into the vector, whilst the pGEM-T Easy system uses a ligase. For both kits, cloning reactions were carried out according to the manufacturers instructions and in both cases the cloning reaction was left for the longest suggested time (1 hr at room temperature for the TOPO TA cloning system and overnight at 4ºC for the pGEM-T Easy vector system). Two different systems were used as recovering transformants with the correct insert proved problematic.

The products of the cloning reaction were transformed into the TOPO TA cloning TOP10 chemically competent *Escherichia coli* cells (Cat. No. K4600-40), TOPO TA cloning TOP10F' chemically competent *E. coli* cells (Cat. No. K4650-40), or New England Biolabs NEB 5-alpha competent *E. coli* cells (Cat. No. C2991H) depending on which cells were available. Transformations were carried out according to the

manufacturers instructions. 2 µl of cloning reaction product were used for all transformations. Heat-shocking was carried out in a water bath.

Cells were screened for the presence of a plasmid with an insert. Transformed cells were plated onto LB nutrient agar plates (7.5 g agar per 500 ml LB) containing carbenicillin (60 µg/ml). Plates were prepared by plating 40 µl X-gal (20 mg/ml in dimethlyformamide) and if TOP10F' or NEB 5-alpha cells had been used for the transformation 10 µl 100 mM IPTG. Both the pCR II-TOPO vector and the pGEM-T Easy vector contain an ampicillin resistance gene allowing only transformed cells to grow in the presence of ampicillin or its derivatives (such as carbenicillin). Both vectors also have their insert site within the coding sequence of ß-galactosidase. When grown in the presence of X-gal cells with an insert have a disrupted ß-galactosidase and appear white, whilst cells without an insert have a functional ß-galactosidase and appear blue.

Colonies with an insert were picked using a 10 µl pipette tip or a sterile toothpick and spotted onto an LB agar plate containing carbenicillin (60 µg/ml) and colonies were grown overnight at 37ºC. To confirm whether the insert was of the expected size, colony PCR was performed on the colonies.

2.1.6 Colony PCR

Colony PCR was carried out using the Roche Taq DNA Polymerase, the AB gene dNTP set (high concentration) and with primers designed to bind to the SP6 and T7 polymerase sites flanking the insert ordered from Thermo Electron or MWG. dNTP and primer stocks were diluted to the given concentration with Milli-Q water. Reactions were carried out in a total volume of 25 µl with the following volumes of reagents:

2.5 μl 10x buffer
21 μl Milli-Q water
1.0 μl dNTP (5mM)
0.2 μl SP6 primer (100nM)
0.2 μl T7 primer (100nM)
0.1 μl Taq DNA polymerase

Colonies were transferred into the reaction mix using a 10 μl pipette tip or a sterile toothpick. 6 -12 colonies were picked for each cloning reaction to increase the chances of picking a colony with the correct insert. Colony PCR was carried out in a thermocycler using a PCR cycle consisting of 1 cycle extended DNA denaturation of 2 min at 94ºC, followed by 35 cycles of 30 sec denaturation at 94ºC, 45 sec annealing at 50ºC and 1 min extension at 72ºC, followed by a final extension step of 7 min at 72ºC.

Colonies containing the correct sized insert were cultured to amplify the number of cells with the insert. Colonies were picked with a 10 μl pipette tip or a sterile toothpick and transferred to culture tubes containing 1 ml LB medium containing carbenicillin (60 μg/ml) and grown overnight at 37ºC on a shaker at 200 rpm.

2.1.7 Minipreps

To isolate the plasmid DNA from the bacterial cells, minipreps were performed using the QIAGEN QIAprep Spin Miniprep Kit (Cat. No. 27106) according to the manufacturers instructions. Cells cultures were first transferred to 1.5 ml Eppendorf tubes and cells were pelleted by centrifugation in a microcentrifuge at 8000 rpm. Once resuspended, the kit was used to lyse cells under alkaline conditions after which, DNA was bound to a silica membrane in the presence of high salt and washes performed to remove endonucleases and salts. The plasmid DNA was eluted into 50 μl elution buffer.

2.1.8 Sequencing

Sequencing reactions were carried out using the Applied Biosystems BigDye Terminator v1.1 Cycle Sequencing Kit (Cat. No. 4337450). The manufacturers instructions were followed with minor modification. Reactions were carried out in a total volume of 10 μl with the following volumes of reagents:

2 μl	5x BigDye sequencing buffer
3.5 μl	Milli-Q water
1 μl	sequencing primer (3 nM)
2.5 μl	plasmid
1 μl	BigDye Terminator ready reaction mix

Sequencing primers were designed to bind to the polymerase sites that flank the insert region (primer sequences given in appendix 2, table A2.3), and each insert was sequenced from both ends. Sequencing was carried out in a thermocycler with an initial denaturation step of 1 cycle of 3 min at 96ºC, followed by 25 cycles of 20 sec at 96ºC, 10 sec at 50ºC and 4 min at 60ºC.

Sequencing reactions products were sent to the Natural History Museum Sequencing Facility for the sequence to be read. They were sent as dried DNA pellets. To pellet the DNA, sequencing reaction products were first precipitated by adding 20 μl Milli-Q water, 70 μl 100% ethanol and 2 μl sodium acetate (3 M) and incubating for 1 hr at room temperature. Precipitated DNA was pelleted by centrifugation at 13000 rpm in a microcentrifuge for 20 min. The liquid phase was discarded and the pellet washed by the addition of 100 μl 70% ethanol. The ethanol was removed and the pellet left to dry by placing in a rack on a 50ºC heating block for approximately 15 min.

2.2 Phylogenetic techniques

I now present the methods used for the analysis of pancrustacean phylogeny (chapter 3). These include methods used for constructing the multigene dataset, analysing the phylogenetic signal in the dataset, running the phylogenetic analyses and using decision criteria to compare between different models.

2.2.1 Compiling the dataset – an overview

To compile a large multigene dataset for investigating pancrustacean phylogeny a search of the GenBank database (Benson, *et al*., 2007, http://www.ncbi.nlm.nih.gov) was carried out to collect arthropod sequences representing as many genes as possible. This search recovered a range of genes from the different datasets previously used to analyse arthropod phylogeny. The genes with the broadest representation of arthropod sequences were the nuclear ribosomal RNAs 18S and 28S, the small nuclear RNA U2, the nuclear protein coding genes elongation factor-1α (EF-1α), RNA polymerase II (PolII), elongation factor-2 (EF-2), histone H3, enolase and glyceraldehyde 3-phosphate dehydrogenase (G3PDH), and complete mitochondrial genomes. The sequences for these different genes did not necessarily represent the same species. However, if the different genes had sequences for species that could be confidently assigned to a monophyletic group, then a composite sequence representing that higher "taxonomic unit" could be used in the phylogenetic analysis.

In order to construct such taxonomic units, the species for which sequences were recovered were first organised into monophyletic groups. The criteria guiding which groups were chosen are outlined in section 2.2.2. For several of these groups, there were gaps in the dataset where sequences were missing for particular genes. Where material could be obtained, genes were sequenced to fill the gaps (see section 2.2.3). Sequences for the different genes within a monophyletic group were then concatenated to give multigene sequences representing the different taxonomic units. This is detailed in section 2.2.4.

2.2.2 Arranging sequences into monophyletic groups

The species for which there was sequence data were arranged into monophyletic groups that represented a diversity of pancrustacean taxa.

The major crustacean groups

Sequences were grouped into the major crustacean subdivisions: the Malacostraca, the Branchiopoda and the recently discovered Cephalocarida and Remipedia. As the maxillopods are now thought to be a polyphyletic assemblage (Mallatt and Giribet, 2006, Regier, *et al.*, 2005, Wills, 1998), sequences were grouped into the major maxillopod taxa, in particular the Cirripedia and the Copepoda. Sequences from other maxillopod taxa that were well represented in GenBank were also included. Also it is unclear whether the two main divisions of the ostracods – the Myodocopa and the Podocopa – form a monophyletic group (Horne, 2005, Regier, *et al.*, 2005), so these were treated as separate clades.

Hexapod taxa and outgroups

As the monophyly of the hexapods has been questioned (Nardi, *et al.*, 2003), it was necessary to distinguish between the true insects (the Insecta) and the entognathous hexapod taxa. Further, as there has been little consensus regarding the position of the different entognathous hexapod groups relative to the insects (Luan, *et al.*, 2005), it was important to distinguish between the Diplura and the Collembola. The Protura are poorly represented in GenBank and so were not considered. As outgroup taxa, both Myriapoda and Chelicerata were used as there has been a large amount of debate as to whether the myriapods are the sister group of the Pancustacea (Mandibulata hypothesis) or whether they are the sister-group of the Chelicerata (Myriochelata/Paradoxopoda hypothesis) (Mallatt, *et al.*, 2004, Pisani, *et al.*, 2004). A range of outgroup taxa were used as it has been shown that outgroup choice can affect the results of a phylogenetic analysis (Rota-Stabelli and Telford, 2008).

Subdividing groups

Where possible, attempts were made to break these major groups into smaller subgroups, as this would provide more sequences, representing the diversity within the groups. For example, rather than compiling a single chimeric sequence representing the Malacostraca, attempts were made to group the different malacostracan sequences according to the five major subdivisions of the Malacostraca, namely the Eucarida, Peracarida, Hoplocarida, Syncarida and Phyllocarida. This often introduced gaps into the final chimeric sequence.

2.2.3 Sequencing additional genes

Complete 28S ribosomal RNA sequences were added for four taxa: *Porcellio scaber* (Malacostraca, Peracarida), *Folsomia candida* (Collembola), *Lepas* sp. (Cirripedia) and *Calanus simullimus* (Copepoda). Partial 28S was added for *Balanus crenatus* (Cirripedia). Genomic DNA samples were kindly donated for *Lepas*, and *B. crenatus* by Prof. Jean Deutsch and for *C. simullimus* by Dr Charles Cook, and RNA for *F. candida* was kindly donated by Dr Sarah Bourlat.

Genomic DNA extraction

P. scaber specimens were collected from a London garden and identified using a number of online keys. To minimise food contamination animals were starved for a week in a Petri dish with damp tissue paper. The animals were killed by immersion in 100% ethanol and a single specimen was ground in liquid nitrogen with a pestle and mortar. Genomic DNA was extracted from this specimen using the QIAGEN Genomic-tip 20/G kit (Cat. No. 10223) (the mass of the animal fell below the 20 mg cutoff that the kit is suitable for) in conjunction with the QIAGEN Genomic DNA Buffer Set (Cat. No.19060). The kit was used to lyse cells before binding genomic DNA to an anion-exchange resin under low-salt and pH conditions, whilst impurities were washed off. DNA was then eluted in a high-salt buffer before being precipitated to remove salts.

DNA was resuspended in 0.1 ml pH 8.0 TE buffer. To confirm that the extraction was successful, 2 μl samples were run on a 1% agarose gel.

Polymerase chain reaction

The 28S gene was amplified by a standard PCR protocol, as described in section 2.1.1 (again using a standard Taq DNA polymerase), with 1 cycle extended DNA denaturation of 2 min at 94ºC, followed by 35 cycles of 30 sec denaturation at 94ºC, 30 sec annealing at a temperature as calculated in section 2.1.1 and 2 min extension at 72ºC, followed by a final extension step of 10 min at 72ºC. Degenerate primers designed against an alignment of metazoan taxa at a range of sites along the 28S gene, were kindly provided by Dr Sarah Bourlat (primer sequences are given in appendix 2, table A2.1). As 28S is a large gene (approximately 4 kb long), it proved difficult to amplify the whole gene as one fragment (the *Taq* DNA polymerase used was suitable for templates ≤ 3 kb – see section 2.1.1). Therefore, primer pairs were chosen that would allow 28S to be amplified in smaller (typically about 1.5 kb) overlapping fragments. In certain cases, there appeared to be difficulties in the subsequent cloning of the PCR products. As a potential problem was the size of the fragments, in certain cases, primer pairs were chosen to amplify smaller fragments (approximately 0.5 – 1 kb).

Gene cloning and sequencing

Amplified fragments were cloned and sequenced as described in section 2.1. In some cases, the insert was large and so sequencing from the SP6 and T7 primers did not produce a long enough sequence to cover the whole insert. In these situations sequencing was repeated using primer sites that lay within the insert. Either the original 28S primers were used if there was an appropriately positioned site, or new primers were designed, which were complementary to sequence in the insert. To confirm that the sequences were from the required gene from the specimen of interest, a BLAST (Basic Local Alignment Search Tool, Altschul, *et al.*, 1990) search was performed. The BLAST algorithm compares a query sequence against sequences in a database and gives statistically supported alignments. Before the BLAST search, the primer sequence was

identified and removed from the sequence file, as was any plasmid sequence. The sequences were compared against the GenBank database using the blastn algorithm, which compares a nucleotide sequence against a nucleotide database. If the highest scoring hits were from the 28S of other closely related arthropods, then it was most probable that the required fragment had been sequenced.

Sequence assembly

The complete 28S sequences were assembled from the smaller overlapping fragments using the SeqMan software from DNASTAR Lasergene v7.0. Sequences were imported and assembled into a contig. Primer and vector sequences were removed, and the chromatogram was inspected and the fragment sequences were truncated where the quality of the peaks deteriorated. Where possible, multiple clones were sequenced for each fragment (ideally three or four), and the sequences were inspected for ambiguous sites, where bases differed between the fragments. Where these ambiguities were present, the chromatogram was checked to see if either sequence had any obvious anomalies, otherwise, the site was marked with ambiguous nucleotide characters.

2.2.4 Constructing concatenated sequences

Selecting sequences and taxa for concatenation

Once the additional sequences had been added, genes that did not have a broad representation across the clades of interest (described in section 2.2.2) were discarded; namely the snRNA U2, the nuclear protein coding genes enolase and G3PDH. This left the rRNAs 18S and 28S, EF-1α, PolII, EF-2, H3 and the complete mitochondrial genomes. The sequences of these genes for the different species within each monophyletic group were then combined to give a multigene sequence representing the group. The accession numbers for these sequences are given in appendix 1, table A1.1.

There is always a risk that sequences from taxa even within the same monophyletic group may evolve under different pressures and at different rates. Therefore, attempts

were made not to concatenate sequences from phylogenetically distant taxa, if it could be avoided. Within each clade, sequences were selected for concatenation from species that formed as small a monophyletic group as possible, provided that this did not lead to a loss of large amounts of sequence data. If decreasing the size of the taxonomic group only led to the loss of a small amount of sequence data (typically fewer than 400 extra sites out of a total alignment length of near 17000) then the smaller group was used.

For example, within the malacostracan subgroup Peracarida, the three taxa *Armadillium valgare*, *Asellus aquatica* and *Porcellio scaber* representing the order Isopoda, together provided sequences for 18S, 28S, EF-1α, Pol II, EF-2 and H3 comprising around 9600 bases. *Armadillium* and *Porcellio* belong to the suborder Oniscidea whilst *Asellus* belongs to the suborder Asellota. Removing *Asellus* would decrease the total length of the concatenated sequence by approximately only 300 bases (by removing the H3 sequence). However, this would also decrease the size of the taxonomic group over which sequences would be concatenated from an order (Isopoda) to a suborder (Oniscidea) (taxonomic rankings from Martin and Davis, 2001). Therefore, the sequences from *Asellus* were removed to maintain the smaller taxonomic group. This process resulted in a range of taxonomic units representing a range of differently sized taxonomic groups.

Sequence alignment

Once all the genes had been chosen and the taxonomic units defined, sequences were aligned for each gene in the dataset. For protein coding genes, nucleotide sequences were aligned according to their translated amino acid sequence using the TranslatorX software (Telford, unpublished).

For the rRNAs, sequences were aligned to include secondary structural information. 28S and 18S rRNA sequences aligned according to their secondary structure were downloaded from the European Ribosomal RNA database (http://bioinformatics.psb.ugent.be/webtools/rRNA) in the dedicated comparative sequence editor (DCSE) format. These were converted into nexus format using the Ystem software (Telford, *et al.*, 2005) and used as a template for the alignment of the

arthropod rRNA sequences, using the profile alignment mode in ClustalX (1.83). For 28S, the sequences were aligned to the five ecdysozoan taxa present in the original DCSE file, and for 18S, the eleven arthropod taxa were used (sequences from closely related taxa, such as the several Diptera, were removed). The taxa used and the accession numbers for the sequences are given in appendix 1, table A1.2.

The Xstem and Ystem software (Telford, *et al.*, 2005) were used to convert the secondary structure information in the DCSE files into a form that could be used by phylogeny software such as MrBayes. The quorum values for Ystem were set so that for a site to be annotated as a stem site, it had to be present in 3/4 of the annotated taxa.

All alignments were inspected by eye in MacClade 4.06 and areas of poor alignment were realigned manually. Sites that could not be aligned satisfactorily across taxa were excluded, and sites within single taxa that could not be aligned convincingly were replaced with "?". At this point the mitochondrial rRNAs (12S and 16S) and the mitochondrial protein coding gene ATP8 were removed from the dataset as they contained too few sites that could be aligned convincingly.

Concatenating to produce chimeric sequences

The aligned sequences of the different genes within each taxonomic unit were concatenated using a Perl script for assembling chimeric sequences from expressed sequence tags (see Bourlat, *et al.*, 2006). To produce the multigene sequences for some taxonomic units, a choice had to be made between several species with sequences for the same gene. For example, within the malacostracan taxon Oniscidea, a choice had to be made between 18S from *Armadillium vulgare* and from *Porcellio scaber*. The Perl script ranked all available sequences for that taxon according to their average distance from all other sequences in the alignment. A composite was built up using as much of the shortest average distance sequence as was present and then adding to it missing regions (if any) from the next ranked sequence until as full a length sequence as possible was built up. This was also used to deal with EF-1α in *Drosophila*. Here there appeared to be two paralogues for the gene, so both copies were put in the alignment and the Perl script was used to select the shortest branch.

Further preparation

In the mitochondrial genome of arthropods, the codon AGG has been shown to be variable in the amino acid it codes for (Abascal, *et al*., 2006). As this could be a source of homoplasy and bias, the codon was replaced by NNN, or X when sequences were coded as amino acids. For phylogenetic analyses where protein coding genes were coded as amino acids, the nucleotide sequences were translated to amino acid sequences using MacClade 4.06.

2.2.5 Analyses of signal in the dataset

Three methods were used to examine sets of aligned sequences for their phylogenetic content (likelihood-mapping and saturation plots) and for their compositional homogeneity (nucleotide composition plots).

Likelihood-mapping

Likelihood-mapping (Strimmer and vonHaeseler, 1997) implemented in TREE-PUZZLE 5.2, is a method to visualise the phylogenetic content of a set of aligned sequences. Quartets of taxa are sampled, and for each quartet the likelihood of each of the three possible fully resolved topologies is calculated. The more signal in the data, the more quartets where one topology is much more likely than the other two. This is represented graphically by plotting the likelihood of quartets on a triangular plot where points in the corners represent fully resolved quartets, points in the lateral regions represent partly unresolved quartets and points in the centre represent fully unresolved quartets. The proportion of fully resolved quartets can be taken as a measure of phylogenetic signal in the data. The analysis was run modelling substitutions using a General Time Reversible (GTR) model (where each different nucleotide substitution can occur at a different rate) with a four-category gamma distribution (to avoid problems of underparameterisation). For other settings, the defaults were used. Taxa for which there was over 90% missing sequence were excluded from the analysis.

Saturation plots

Saturation plots are a method to visualise whether a set of aligned sequences has been saturated with mutations, which would obscure any phylogenetic signal. A number of different variants of this type of plot have been described in the literature; I follow the method used by Negrisolo *et al*. (2004). For each taxon pair, the uncorrected ("p") distance calculated in PAUP 4.0b10, was plotted against the maximum likelihood distance calculated by TREE-PUZZLE 5.2 (see above). Where signal is present, the uncorrected ("p") distance increases linearly with the maximum likelihood distance. However, as the uncorrected ("p") distance does not account for base pairs where multiple substitutions have occurred, when the signal is saturated the plot levels off at an uncorrected ("p") distance of 0.75.

Nucleotide composition plots

Compositional heterogeneity within a dataset can be problematic for phylogenetic reconstruction. For example, shared compositional biases can often lead to artefactual attraction between taxa (Hassanin, *et al*., 2005). Nucleotide composition plots allow a simple visual comparison of the nucleotide composition between taxa. The proportion of each base was calculated using the *show nucleotide frequencies* option in MacClade 4.06 and these were plotted for each taxon using Microsoft Excel v.X. Where nucleotide composition was largely homogeneous amongst taxa, the plots of nucleotide frequencies appear flat for each base. Where nucleotide composition was heterogeneous, nucleotide frequencies appear more variable between taxa.

2.2.6 Bayesian phylogenetic analysis

Bayesian phylogenetic analyses were performed using MrBayes v3.1.2 (Huelsenbeck and Ronquist, 2001, Ronquist and Huelsenbeck, 2003). The dataset was partitioned differently depending on the modelling strategy used for each analysis, and the substitution model for each partition was as specified in the modelling strategy. For the analyses where the protein coding genes were coded as amino acids, model-jumping

between fixed rate amino acid models was implemented, allowing the MCMC to pick the best model. The Cprev model of amino acid substitution was replaced by a model of amino acid substitution based on Metazoan genomes (Rota-Stabelli and Telford, unpublished). The substitution matrix, transition-transversion ratio, shape of the gamma distribution and state frequency parameters were unlinked between the different partitions. Constraints on tree topology were only used when stated. For other settings the default conditions were used.

The analysis was run for an initial 4 million generations. Four chains were run, one cold and three heated (using the default settings), and the posterior distribution was sampled every 100 generations. Log likelihoods were plotted against generation for every 1000th generation and inspected by eye. If the distribution appeared to have reached a plateau (taken as showing no obvious upwards trend for over 2 million of the 4 million generations), the analysis was terminated. If a plateau had not been reached, the analysis was successively run for a further 2 million generations until a plateau appeared to have been reached. Once the log likelihood values were judged to have reached a plateau, a consensus tree was generated using the *sumt* command in MrBayes. The burnin was set to exclude all but the final 2 million generations.

2.2.7 Calculating convergence diagnostics

Two diagnostics were calculated to examine whether two runs of a modelling strategy had converged on the same posterior probability distribution: the average standard deviation of taxon bipartiton posterior probabilities – referred to as the split frequencies – which indicates the extent of topological convergence between the two runs, and the distribution of log likelihoods, which indicates whether the two runs have reached a plateau at the same distribution of log likelihood values.

Calculating split frequencies

Using the online software AWTY (Nylander, *et al*., 2008), the frequency of each taxon bipartition in the posterior sample for each run was output using the *Showsplits* analysis.

To assess the difference in the frequency of each bipartition, the standard deviation was calculated for each bipartition (the standard deviation was used rather than the difference, as this is a standard approach used in software such as MrBayes for assessing convergence). The average standard deviation of taxon bipartition frequencies was then calculated as an overall diagnostic for the similarity in the sample posterior of the two runs. The smaller the average, the better the convergence.

Distribution of log likelihoods

The arithmetic means of the post burnin log likelihoods were calculated using Microsoft Excel v.X. To make computation easier, these were calculated for every 1000th generation. Additionally, the spread was compared by calculating the range in which 95% of likelihood values fell. Again this was calculated in Microsoft Excel v.X using the *PERCENTILE* option. These two measures were plotted for each pair of runs to allow graphical comparison.

2.2.8 Bayes factors

Bayes factors test whether the data supports one of two competing models. The hypotheses could be different models for sequence evolution (such as different partitioning strategies) or different phylogenetic hypotheses, where the Bayesian analysis was run with a constrained tree topology. The Bayes factor (BF) is the ratio of marginal likelihoods (the likelihood of the data under a particular model after integrating across parameter values) from two competing models (Brown and Lemmon, 2007). Bayes factors are not used in a normal statistical test, where a hypothesis is accepted or rejected relative to some arbitrary cutoff; rather they evaluate the relative merits of competing models (Nylander, *et al.*, 2004).

The test statistic $2\ln(BF_{21})$ was used, where BF_{21} is the Bayes factor for model 2 over model 1, and this was interpreted according to the guidelines of Kass and Raftery (1995 cited in Nylander, *et al.*, 2004) (see table 2.1). The value of 2lnBF = 10 is often used alone as a simple cutoff, as in Brown and Lemmon (2007), where $2\ln(BF_{21}) > 10$

indicates significant support for model 2, $10 > 2\ln(BF_{21}) > -10$ indicates ambiguity and $2\ln(BF_{21}) < -10$ indicates significant support for model 1.

Table 2.1. Interpretations of Bayes factors (BF_{21}) based on Nylander *et al.* (2004).

$2\ln(B_{21})$	Evidence against M_1
0 to 2	not worth more than a bare mention
2 to 6	positive
6 to 10	strong
>10	very strong

The marginal likelihood for a particular model is difficult to calculate. However, the harmonic mean of the likelihood values can be used as an estimate (Newton and Raftery, 1994 cited in Brown and Lemmon, 2007). Therefore, the $2\ln(BF_{21})$ statistic was calculated as:

$$2\ln(BF_{21}) = 2[\ln(HM_2) - \ln(HM_1)]$$

where HM_2 is the harmonic mean of the posterior sample of likelihoods from model 2 and HM_1 is the harmonic mean of the posterior sample of likelihoods from model 1. Harmonic means were output by the *sump* command in MrBayes. The same burnin was used as for the consensus trees (section 2.2.6).

2.2.9 Information criteria

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) indicate the fit of a model to the data, taking into account the complexity of the model: the likelihood of the model is penalised as a function of the number of parameters (Posada, 2003). Both criteria should be applied in a likelihood framework; however, here they are used in a Bayesian framework, assuming that the harmonic mean of the posterior distribution is a reasonable estimate for the maximum likelihood (as in McGuire, *et al.*, 2007). The AIC and BIC were calculated as follows:

$$AIC = -2l+2K$$

$$BIC = -2l+K\ln(n)$$

where l is the log likelihood (estimated as the post burnin harmonic mean as used for the Bayes factors), K is the number of estimable parameters and n is the sample size (approximated by the total number of characters in the alignment). A difference in the value of the AIC or BIC (Δ(AIC) or Δ(BIC)) >10 between two models is taken as strong support for one model over another (Posada and Buckley, 2004).

2.3 Embryological techniques

I now present the various methods used for producing the *Tribolium* and *Drosophila* gene expression patterns documented in chapters 4 and 5. I include the methods used for identifying the orthologues of *Drosophila* genes in the *Tribolium* genome.

2.3.1 Stock maintenance

Tribolium

Three vials of *Tribolium castanuem* of the San Bernardino wildtype strain were kindly donated by Dr Gregor Bucher. Beetle stocks were maintained in 1 l Tupperware boxes on wholemeal flour at room temperature. Wholemeal flour was first sterilised by heating at 65ºC for at least 24 hr, and was then passed through a 500 μm sieve to remove large particles to aid subsequent separation of beetles from the flour.

Beetles were transferred onto clean flour every 6 months. Adult beetles, pupae and final instar larvae were separated from the flour using an 800 μm sieve. This sieve also

separated exuviae and dead beetles from the flour. The exuviae and some of the dead beetles were removed by gently blowing over the sieve plate. The adults and larvae were separated from the remaining dead beetles by paper transfer. Beetles were placed onto a sheet of paper which was turned vertically. Only the living beetles held onto the paper allowing their separation from dead individuals. Sieve plates were always sterilised after use by heating to 65ºC.

Drosophila

Five vials of *Drosophila melanogaster* were kindly donated by Prof. Ernst Wimmer. Two of the lines contained the UAS-*ems* construct, whilst three contained the Gal4 driver for use in misexpression the *empty spiracles* gene across the anterior of the embryo (Schöck, *et al.*, 2000). Embryogenesis of all lines was wildtype and the most vigorous line was used for embryo collection. Flies were maintained in vials on ASG food with a few grains of dried yeast at 25ºC and were transferred onto new food approximately every two weeks, using CO_2 to anaesthetise the flies during transfer. The recipes for the various *Drosophila* media are given in table 2.2.

Table 2.2. Media for fly culturing. The recipes given are for the stated amount of the media. Volumes were altered when different amounts of the media were needed.

Medium	Recipe	
ASG Food	Water	500 ml
(makes 500 ml)	Agar	5 g
	Sugar	42.5 g
	Yeast	10 g
	Maize	30 g
	Nipagin	12.5 ml
Grape plates	Water	1000 ml
(makes ~80 purps)	Agar	50 g
	Grape juice	600 ml
	More water	100 ml
	Nipagin	42 ml
Yeast paste	Dried yeast	10 g
(makes 10 ml)	Water	10 ml

2.3.2 Embryo collection

Tribolium

Adult beetles were transferred onto plain flour by paper transfer (as described in section 2.3.1). Plain flour was first sterilised at 65ºC for at least 24 hr and was then passed through a 250 μm sieve. Embryos were collected every 4-5 days; during this period of time at room temperature embryos reached early germband retraction. Beetles were first removed from the flour by passing through a 500 μm sieve. Embryos were then separated from the flour by passing through a 250 μm sieve. Beetles were returned to the flour and left to continue laying. Embryo collection was continued until egg production was low, at which point the beetles were transferred back to wholemeal flour.

Drosophila

Adult *Drosophila* were transferred from vials into bottles containing ASG food and a few grains of dried yeast to allow culture sizes to increase. After 3-4 weeks at 25ºC adult flies were transferred to new bottles containing ASG food and a few grains of dried yeast and allowed to lay eggs. 10 days after laying began, new flies started to emerge. Old adults were removed and newly emerged adults were collected. 50 female flies and 30-40 male flies were placed in new bottles containing ASG food and a few grains of dried yeast and allowed to mate. After 1-2 days, female flies were transferred to grape agar purps with yeast paste and allowed to lay. Purps were replaced twice daily. After one day of laying, embryo collection began. Flies were left to lay for 10-14 hr, after which time purps were collected. Embryos were washed off the purps using PBT (see table 2.3) and transferred into a 15 ml Falcon tube using a plastic Pasteur pipette. Embryo collection was carried out for 2 days. The recipes for the various *Drosophila* media are given in table 2.2.

Table 2.3 Reagents for *Tribolium* and *Drosophila* embryology. The recipes given are for the stated amount of the reagent. Volumes were altered when different amounts of the reagents were needed.

Reagent		Recipe
PBT	250 ml	PBS
(makes 250 ml)	500 µl	10% TWEEN 20
PEMS	400 ml	Water
(makes 400 ml)	12.08 g	Pipes
	800 µl	$MgSO_4$ (1 M)
	800 µl	EDTA (pH 8.0, 0.5 M)
		to pH 6.9 with NaOH
Hybe-A	25 ml	Deionised formamide
(makes 50 ml)	12.5 ml	20x SSC
		to pH 5.5 with HCl
		to 50 ml with water
	1 ml	Sonicated salmon sperm DNA (10 mg/ml)
	250 µl	tRNA (20 mg/ml)
	25 µl	heparin (100 mg/ml)
Hybe-B	25 ml	Deionised formamide
(makes 50 ml)	12.5 ml	20x SSC
		to pH 5.5 with HCl
		to 50 ml with water
Blocking buffer	500 g	Bovine serum albumin
(makes 50 ml)	1 ml	Sheep serum
		to 50 ml with water
Staining buffer	5 ml	Tris-Cl (pH 9.5, 1 M)
(makes 50 ml)	2.5 ml	$MgCl_2$ (1 M)
	1 ml	NaCl (5 M)
		to 50 ml with water
	250 µl	10% TWEEN 20
Inactivation buffer	50 ml	Hybe-B
(makes 50 ml)	500 µl	10% TWEEN 20
	750 µl	20% SDS
	25 µl	Heparin (100 mg/ml)
	50 µl	Sonicated salmon sperm DNA (10 mg/ml)

Note. All reagents made up according to (Sambrook and Russell, 2001) unless stated otherwise. For 10% TWEEN 20 use SIGMA TWEEN 20 (Cat. No. P9416) diluted 10-fold in water, for sheep serum use SIGMA Sheep Serum (Cat. No. S2263), and for bovine serum albumen use SIGMA Albumin from bovine serum (Cat. No. A4503).

2.3.3 Embryo dechorionation

For both *Tribolium* and *Drosophila*, embryos were transferred to egg baskets and dechorionated by washing in 50% bleach for 2 min. Egg baskets were made by attaching a polyamide screen 100 μl mesh over the end of a 50 ml Falcon tube. Embryos were then washed in deionised water to remove bleach.

2.3.4 **Tribolium** *RNA extraction and cDNA synthesis*

Tribolium RNA was extracted from embryos using the TRIZOL extraction protocol kindly provided by Dr Nikola Michael Prpic. Embryos were collected over 5 days as this time period contained the stages of interest (see section 2.2.2), and therefore the relevant mRNAs were being expressed. Dechorionated embryos were transferred into a sterile Eppendorf tube to a depth of 3-5mm and homogenised in 750 μl TRIzol (Invitrogen Cat. No. 15596-026). To remove cell debris, the homogenate was centrifuged for 10 min at 13000 rpm at 4ºC after which the clear pinkish liquid was transferred to a new tube and incubated for 5 min at room temperature. 200 μl chloroform were added and mixed by gently shaking. This mixture was incubated for 10 min at room temperature before centrifugation for 15 min at 4ºC, producing two phases.

RNA was contained in the top phase which was transferred to a new tube, and the RNA precipitated by adding 500 μl isopropanol and mixing by gently shaking. The mixture was incubated for 10 min at room temperature and then centrifuged for 10 min at 4ºC. The supernatant was discarded, leaving a pellet of RNA. This was washed by adding 1 ml 70% ethanol and incubating for 5 min at room temperature, before spinning at 13000 rpm for 10 min at 4ºC. The ethanol was removed and the pellet was left to air dry on ice (to prevent RNA degradation) until all the ethanol had evaporated (approximately 10 min). The pellet was dissolved in 50 μl Milli-Q water. This gave an RNA extract at a concentration of approximately 3000 ng/μl (Coulcher, J. F., personal communication). 2 μl of the RNA extract was run on a 1% agarose to confirm the extraction had worked.

RNA extracts were stored at –80ºC. cDNA was synthesised from the RNA as described in section 2.1.2. cDNA was synthesised in batches of 20 μl and stored as a stock.

2.3.5 Identifying Tribolium *orthologues of* Drosophila *genes*

BLAST search of BeetleBase

Tribolium orthologues of *Drosophila* genes were identified in the *Tribolium* genome by a BLAST search of the *Tribolium* Genome Database resource BeetleBase V2.0 (BeetleBase website: http://www.bioinformatics.ksu.edu/BeetleBase). The protein sequences of the *Drosophila* genes of interest were downloaded from the GenBank database for use as the query sequences in the BLAST search. Using the *BLAST* option in BeetleBase, the query sequences were compared against the *All Tribolium sequences* database using the tblastn algorithm, which compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.

The *All Tribolium sequences* database on BeetleBase V2.0 largely contains unannotated contigs as well as ESTs and published sequences. Therefore, several BLAST searches recovered isolated stretches of sequence similarity within the contigs. The stretches of nucleotide sequences showing similarity to *Drosophila* were extracted. Often successive stretches of alignment were identified in the same contig that matched successive regions of the query sequence. In these cases it was assumed that the *Tribolium* sequences were parts of the same gene, separated either by more divergent areas of poorer alignment with *Drosophila*, or by introns.

Reciprocal BLAST

The stretch of alignment with the highest E-value was likely to belong to the direct orthologue of the original *Drosophila* gene. To confirm that this was the case, the highest scoring alignment was used as the query sequence and a BLAST search was performed against the *Drosophila melanogaster* protein database on NCBI BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi), using the blastx algorithm, which compares a

nucleotide query sequence translated in all reading frames against a protein sequence database. In cases where multiple stretches of alignment appeared to correspond to one gene, a concatenated sequence was used at the query. For a direct one-to-one orthology, the original *Drosophila* gene had to be recovered as the highest scoring sequence, judged by the E-value. Additionally, the next highest scoring stretches of alignment in the *Tribolium* genome were also compared to the *Drosophila* protein database by a BLAST search. These could not recover the original *Drosophila* gene, as this would suggest that there had been *Tribolium* gene duplications or *Drosophila* gene losses.

2.3.6 Cloning Tribolium *orthologues*

Tribolium orthologues were amplified from *Tribolium* cDNA (see section 2.2.4) by PCR (as described in section 2.1.2). A PCR cycle was used consisting of 1 cycle extended DNA denaturation of 2 min at 94ºC, followed by 35 cycles of 1 min denaturation at 94ºC, 2 min annealing at a temperature as calculated in section 2.1.1 and extension for the appropriate length of time at 72ºC, followed by a final extension step of 10 min at 72ºC. Specific primers were designed against the *Tribolium* sequences identified by the reciprocal BLAST search (section 2.2.5). Primer pairs were typically designed to amplfy partial cDNAs of 0.5-1.0 kb.

Various criteria were taken into account when designing primers. Primers were 21 nucleotides long and the melting temperatures of a primer pair were, where possible designed to be within 2-4ºC of each other, as calculated by "the Wallace rule" (section 2.1.1). Additionally, primers were inspected by eye for any obvious sequence that would anneal to themselves, and primer pairs were inspected by eye for any obvious complementary sequences that could dimerise.

PCR products were purified and cloned, as described is section 2.1. To confirm that the desired partial cDNA had been cloned, the insert was sequenced (as decribed in section 2.1.8). The sequencing products were converted into FASTA files and were compared to the *Tribolium* sequences identified in the reciprocal BLAST search, using MacClade 4.06.

2.3.7 **Drosophila *clones***

Cloned complete cDNAs of *Drosophila* genes were ordered from the *Drosophila* Gene Collection Release 3.0 (DGCr3) through Geneservice. cDNAs were sent cloned into either the pOT2, pFLC-1 or pBS vectors and sent transformed into *E. coli* cells streaked out on LB agar containing antibiotic. Clones were treated according to the distributors instructions. Clones were first streaked out on LB agar plates containing either carbenicillin (60 μg/ml) or chloramphenicol (25 μg/ml) to isolate individual colonies. The antibiotic used depended on the vector: the pFLC-1 or pBS vectors contained an ampicillin resistance gene whilst the pOT2 vector contained a chloramphenicol resistance gene. 10 colonies were picked for each clone and grown up on LB agar containing the appropriate antibiotic. Colonies were picked and cultured in LB containing the appropriate antibiotic (at 1 μl per 1 ml LB) and miniprepped as described in section 2.1.7. To confirm the cDNAs were for the correct genes, the inserts were sequenced as described in section 2.1.8, and the sequence outputs were aligned to sequences of the complete cDNA downloaded from the GenBank database. For the pFLC-1 and pBS vectors, primers complementary to the T7 and T3 polymerase sites flanking the cDNA insert were used for sequencing. For the pOT2 vector, primers complementary to the SP6 and T7 polymerase sites flanking the cDNA insert were used for sequencing

2.3.8 RNA probe synthesis

Labelled RNA probes were synthesised for *in situ* hybridisation. Digoxigenin (DIG) labelled probes were used in standard *in situ* hybridisations for the detection of the transcripts of a single gene. Fluorescein labelled probes were used for detecting the transcripts of a second gene in double *in situ* hybridisations. Probes were synthesised using a protocol provided by Dr Nikola Michael Prpic.

PCR to generate probe synthesis template

One clone with the required cDNA insert was chosen for each gene for probe synthesis. Before transcription, the stretch of the plasmid containing both the insert and the polymerase start sites was amplified by PCR. Reactions were carried out in a total reaction volume of 100 μl, using the Roche Taq DNA Polymerase and the AB gene dNTP set (high concentration) (as described in section 2.1.1). Primers designed to regions flanking the polymerase sites were used (primer sequences are given in appendix 2, table A2.4). Reactions were carried out with the following volumes of reagents:

10 μl	10x buffer
76 μl	Water
4 μl	dNTP (5 mM)
5 μl	Forward primer (10 nM)
5 μl	Reverse primer (10 nM)
0.2 μl	Template
0.5 μl	Taq DNA polymerase

A PCR cycle was used consisting of 1 cycle extended DNA denaturation of 1 min at 94ºC, followed by 30 cycles of 30 sec denaturation at 94ºC, 45 sec annealing 45ºC and 1 min 30 sec extension at 72ºC, followed by a final extension step of 7 min at 72ºC.

PCR products were purified using the QIAGEN QIAquick PCR Purification Kit (Cat No. 28106). Gel purification (section 2.1.3) was not used as the PCR reaction used specific primers against a plasmid vector so it was very unlikely that undesired fragments would be amplified. DNA was adsorbed to a silica membrane in the presence of high concentrations of salt, whilst impurities and contaminants pass through the membrane. DNA is then eluted into an elution buffer. Accoring to the kit manual, the spin columns can bind up to 10 μg DNA, which is eluted into 30 μl elution buffer at a 90-95% efficiency. Therefore, a template concentration of up to approximately 300 ng/μl could be recovered, depending on the efficiency of the PCR reaction.

Probe synthesis

In order to bind to the target mRNA, probes had to be synthesised from the antisense strand. The orientation of the insert in the vector was judged from the sequencing products and the correct polymerase to synthesise the antisense strand was chosen. The transcription reactions were carried out in a total volume of 10 μl using either the Roche SP6 RNA Polymerase set (SP6 RNA polymerase and 10x transcription buffer; Cat. No. 10 810 274 001), the Roche T7 RNA Polymerase set (T7 RNA polymerase and 10x transcription buffer; Cat. No. 10 881 767 001) or the Roche T3 RNA Polymerase set (T3 RNA polymerase and 10x transcription buffer; Cat. No. 11 031 163 001) depending on the clone. Either the Roche DIG RNA labelling mix (Cat. No. 11 277 073 910) or the Roche Fluorescein RNA labelling mix (Cat. No. 11 685 619 910) was used depending on the probe being made. Roche Protector RNase inhibitor was also used to prevent RNA degradation. Reagents were used in the following volumes:

6 μl	Purified PCR product (approximate concentration 300 ng/μl)
1 μl	10x transcription buffer
1 μl	10x RNA labelling mix
1 μl	RNase inhibitor
1 μl	RNA polymerase

The transcription reaction was run for 2 hr at 37ºC in a hybridisation oven.

To terminate transcription 1 μl EDTA (pH 8.0, 0.5 M), 90 μl Milli-Q and 1 μl tRNA (20 mg/ml) were added, and the mixture was gently mixed and spun down. To precipitate the labelled RNA probes 45 μl ammonium acetate (10 M) and 435 μl 100% ethanol were added, the mixture was gently mixed and spun down and then incubated for 1 hr at –20ºC. To wash the probe the mixture was then centrifuged for 20 min at 13000 rpm before removing the liquid and adding 500 μl 75% ethanol. This was incubated on ice for 5-10 min before being centrifuged for 10 min at 13000 rpm. The ethanol was removed and the pellets air dried on ice until residual ethanol had evaporated (approximately 10-15 min). Probes were dissolved in 100 μl Milli-Q water. 2 μl of the probe was run on a 1% agarose gel to confirm synthesis had been successful.

A regular agarose gel (as described is section 2.1.4) was used rather than an RNA denaturing gel. Probes were stored at –80ºC.

Probe concentrations varied (as indicated the intensity of the bands on the agarose gel). However, the manufacturer's instructions for the DIG and Fluorescein RNA labelling mixes state that approximately 10 μg of full length labelled RNA are transcribed from 1 μg linear template DNA. Given the approximate mass of the template DNA of a little under 2 μg (6 μl at approximately 300 ng/μl), approximately 20 μg labelled probe was synthesised, giving a final concentration of around 200 ng/μl (after dissolving in 100 μl).

2.3.9 Embryo fixation

Tribolium and *Drosophila* embryos were fixed according to the protocol of Dr Gregor Bucher. Using a paintbrush, freshly dechorionated embryos were transferred into 30 ml bottles (Fisher, Catalogue number FB73250) containing 12 ml heptane, 4 ml PEMS (see table 2.3) and 600 μl SIGMA 37% formaldehyde (F-1635). Embryos were fixed at room temperature for 30 min on a shaking platform (approximately 200 rpm). Shaking at 37ºC for 20 min did not appear to affect fixation.

Embryos were devitellinised by methanol shocking. 16 ml methanol (room temperature) was added to the fixation mixture and vigorously shaken for 30 sec. The bottle was swirled and devitellinised embryos fell to the bottom whilst embryos with the vitellin membrane attached remained at the water-heptane interface. Devitellinised embryos were collected with a glass pipette and transferred to a 15 ml Falcon tube.

For *Tribolium*, methanol shocking alone often did not recover many embryos. Many embryos with the vitellin membrane attached remained at the water-heptane interface. To remove the vitellin membrane these embryos were repeatedly aspirated and expelled vigorously through a 0.7 mm needle using a syringe (30 or 50 ml). As with the methanol shock, devitellinised embryos fell to the bottom of the bottle and were

collected with a glass pipette. Once collected, embryos were rinsed twice by replacing the methanol and stored in methanol at –20ºC until required.

2.3.10 **In situ *hybridisation***

In situ hybridisation in both *Tribolium* and *Drosophila* was carried out according to the protocol of Dr Gregor Bucher (as described in Wohlfrom, *et al*., 2006). Compositions and recipes for buffers are given in table 2.3. In the following descriptions, rinsing embryos refers to simply replacing the buffer and washing embryos refers to replacing the buffer and rotating on a wheel for a given amount of time.

Embryo preparation

Embryos were removed from storage at –20ºC and transferred into 1.5 ml Eppendorf tubes. A depth of 2-3 mm of embryos in an Eppendorf tube (approximately 200 embryos for *Tribolium* and 300 embryos for *Drosophila*) was sufficient for one *in situ* hybridisation. Embryos were first rinsed in clean methanol and then in 50% methanol/PBT. This was followed by post-fixation in 1 ml PBT with 140 μl 37% formaldehyde for 15 min on a wheel. Embryos were then washed by rinsing twice in PBT followed by three 5 min washes in PBT. This was followed by a Proteinase K digestion; embryos were incubated on a wheel for 5 min in 1 ml PBT with 5 μl Proteinase K (Roche Proteinase K (Cat. No. 3 115 828) diluted ten fold in PBT). The Proteinase K digestion was stopped by rinsing twice in PBT, followed by post-fixing in 1 ml PBT with 140 μl 37% formaldehyde for 15 min rotating on a wheel. Embryos were then rinsed twice in PBT, followed by a 5 min wash in PBT and a further rinse.

Hybridisation

Before hybridisation, embryos were first washed in 250 μl PBT with 250 μl Hybe-B buffer, which was replaced by 250 μl Hybe-B. This was then replaced with 250 μl Hybe-A buffer, and the embryos were prehybridised for 1 hr at 65ºC in a water bath. After prehybridisation as much Hybe-A was aspirated as possible, the probe was then

diluted in 30 μl which was added to the embryos. The concentrations of the probes varied (as described in section 2.3.8); for details on optimising the volume of probe used see section 2.2.12. Embryos were hybridised overnight at 65ºC.

Post-hybridisation

After hybridisation, 500 μl Hybe-B (prewarmed to room temperature) was added to the embryos, keeping at 65ºC until the embryos settled. This was replaced with 500 μl Hybe-B and incubated at 65ºC for 15 min. Embryos were then transferred to room temperature and diluted by adding 500 μl PBT, after which they were blocked by rinsing and then washing for 5 min, 15 min and then 20 min in 1 ml blocking buffer (the blocking buffer was kept on ice).

The DIG-labelled probe was detected with an alkaline phosphatase conjugated anti-DIG antibody. Embryos were rotated on a wheel for 1 hr in 1 ml blocking buffer with 0.5 μl Roche Anti-Digoxigenin-AP, Fab fragments (Cat. No. 11 093 274 910). After the antibody incubation, embryos were first rinsed twice, and then washed for 5 min, 20 min then twice for 30 min in blocking buffer if still available, otherwise in PBT.

Staining

Expression patterns were visualised using the NBT/BCIP system. The yellow substrate BCIP is metabolised by the alkaline phosphatase coupled to the anti-DIG antibody in the presence of NBT to give a dark blue product. Embryos were first rinsed then washed for 5 min in staining buffer. Embryos were then stained in 1 ml staining buffer with 20 μl Roche NBT/BCIP stock solution (Cat. No. 1 681 451). Embryos were transferred to watch glasses for the stains to develop. Stains were developed in the dark at room temperature. To monitor the development of the stain, embryos were periodically inspected under a dissecting microscope. When the stain had developed to the desired level, the staining reaction was terminated by washing three times for 10 min in PBT. Stained embryos were stored at 4ºC in 1 ml PBT with 100 μl 37% formaldehyde.

2.3.11 Double in situ *hybridisation*

Double *in situ* hybridisation was carried out according to the protocol of Dr Gregor Bucher (as described in Wohlfrom, *et al.*, 2006). The protocol was largely the same as for the standard single *in situ* hybridisation, with a few modifications. The hybridisation step was carried out with probes for the two genes of interest; one DIG-labelled, the other fluorescein-labelled. The expression patters for the two genes were the visualised successively. First, the fluorescein-labelled probe was detected with alkaline phosphatase conjugated anti-fluorescein antibodies and visualised using the Fast Red system. The substrate Fast Red TR/Naphthol AS-MX is metabolised by the alkaline phosphatase coupled to the anti-fluorescein antibody to give an intense red stain. After developing the first stain, the DIG-labelled probes were visualized as for a standard single *in situ* hybridisation. Generally the weaker probe was DIG labeled. The modifications to the protocol will now be detailed.

Embryo preparation and hybridisation

The preparation of embryos was the identical to the standard single *in situ* hybridisation. For the hybridisation step, where the single DIG-labelled probe was added to 30 μl Hybe-A and incubated for 10 min for the standard *in situ* hybridisation, both the DIG-labelled and fluorescein-labelled probes were added for double *in situ* hybridisation.

Visualising the fluorescein-labelled probe

Post-hybridisation washes were carried out as for the standard single *in situ* hybridisation, except that for the antibody incubation, an anti-fluorescein labelled antibody was used rather than an anti-DIG antibody. Embryos were rotated on a wheel for 1 hr in 1 ml blocking buffer with 0.5 μl Roche Anti-Fluorescein-AP, Fab fragments (Cat. No. 11 426 338 910). Subsequent wash steps were as for the single *in situ* hybridisation.

The expression pattern for the fluorescein labelled probe was visualised, using the Fast Red system. Embryos were rinsed and washed for 5 min in 0.1 M Tris-HCl, pH 8.2.

Embryos were stained using Sigma SIGMAFAST Fast Red TR/Naphthol AS-MX Tablets (Cat. No. F4648) in a volume of 1 ml as according to the manufacturers instructions. Stains were developed and monitored as for the NBT/BCIP system. Stains often developed more slowly using the fluorescein probes with Fast Red, so if embryos had not stained to the desired level in 4-5 hr, they were left overnight at 4ºC. When the embryos had stained sufficiently, staining was terminated as for the NBT/BCIP system, except that formaldehyde was not added.

Visualising the DIG-labelled probe

To visualise the DIG-labelled probe, activity of the alkaline phosphatase coupled to the anti-fluorescein antibody had to be inactivated. PBT was replaced by 1 ml inactivation buffer, prewarmed to 65ºC, and incubated at 65ºC for 15 min in a heating block. Embryos were then cooled to room temperature in the inactivation buffer (approximately 20 min). 500 μl inactivation buffer was replaced with 500 μl PBT. From this point embryos were treated as for post-hybridisation in a standard single *in situ* hybridisation, using an alkaline phosphatase couple anti-DIG antibody, and developing the stain with the NBT/BCIP system.

2.3.12 Reducing background

For some genes there were high levels of background. Two steps were varied in an attempt to reduce background: probe concentration, and using a preabsorbed antibody.

Probe concentration

Reducing the amount of probe could reduce background although too little probe could result in a weak signal. As a starting point, *in situ* hybridisation was carried out using 5 μl of probe. If there was a poor signal to background ratio, *in situ* hybridisation was repeated using up to a 50x dilution of the probe concentration.

Antibody preabsorption

The use of a preabsorbed antibody to detect the probe appeared to reduce background levels for several probes. This was particularly noticeable when using fluorescein labelled probes (the effect was less noticeable for DIG labelled probes). To preabsorb the antibody, fixed embryos were transferred to an Eppendorf tube (a similar amount of embryos to an *in situ* hybridisation). Embryos were washed four times for 20 min in blocking buffer. After the last wash, the blocking buffer was removed and replaced with 1 ml fresh blocking buffer and 50 μl undiluted antibody, and this was incubated overnight at 4ºC. The supernatant was removed and stored at 4ºC. For the antibody incubation, to maintain the same antibody concentration, 10 μl preabsorbed antibody was added to the 1 ml blocking buffer.

2.3.13 Embryo preparation and image acquisition

For both *Tribolium* and *Drosophila*, embryos were transferred to a watch-glass and examined under a dissecting microscope. The desired embryos were selected and transferred to a spot of glycerol on a microscope slide using a 10 μl Gilson pipette. For lateral view, a cover slip was placed over the embryo, and the specimen rolled into the correct orientation. For *Tribolium* embryos the head was first freed of yolk using a flame sharpened tungsten needle, wax mounted in a 20 μl pipette tip. For flat mounted *Tribolium* embryos, the embryo was transferred to glycerol and yolk removed, first being broken up with a pair of forceps and then carefully removed with an eyebrow hair, wax mounted in a 20 μl pipette tip. For flat mounted *Drosophila* embryos, the embryos were split dorsally using a flame sharpened tungsten needle, allowing the germband to be flattened. Brightfield and DIC images were taken with a Zeiss AxioImager.M1 coupled to a Zeiss AxioCam HRc. Brightness and contrast were adjusted with the GNU Image Manipulation Program (GIMP) 2.2.10.

Chapter 3: Pancrustacean phylogeny and the position of the insects

3.1 Summary

In this chapter I address the issue of pancrustacean phylogeny and the position of the insects. A number of recent molecular phylogenetic analyses of the arthropods based on a variety of datasets have had a broad enough sampling of crustacean taxa to allow an accurate placement of the insects within the group. However, these different analyses have not reached a consensus on the phylogeny of the Pancrustacea or which crustacean group is the sister taxon to the insects. I have addressed these questions with a multigene Bayesian phylogenetic analysis. I have compiled the various genes used in previous analyses of arthropod phylogeny to produce the largest dataset yet used for a Bayesian analysis of a taxonomically diverse pancrustacean phylogeny, constructing the dataset to give a broad representation of crustacean taxa, and adding new sequences of 28S ribosomal RNA to fill in important gaps. I ran a number of analyses addressing areas of uncertainty in how to model a multigene dataset. First I addressed the heterogeneity in the evolutionary process between the different codon positions of protein coding genes. I then addressed the effect of changing the model of nucleotide substitution. I also investigated the effect of modelling the data with protein coding genes coded as amino acid sequences. These different analyses find strong support for grouping the insects with the other hexapod taxa, and grouping these hexapods with the branchiopod crustaceans. Finally I ran a series of hypothesis tests using Bayes factors

which illustrate that a grouping of the hexapods with the branchiopods receives very strong support over almost all other placements of the hexapods in the Pancrustacea.

3.2 Introduction

As we have seen, the recognition of a crustacean origin for the insects has revolutionised our understanding of insect bodyplan evolution. It has opened up a number of questions, perhaps the biggest being how the insect bodyplan fits into the diversity of crustacean bodyplans that were introduced in chapter 1. Resolving this issue will not only help clarify what transitions took place during the evolution of the insect bodyplan, but will also provide a framework within which to infer the developmental changes that took place during these character transitions. However, as will now be seen, resolving pancrustacean phylogeny and the position of the insects has been problematic. This is the issue that I address in this chapter.

3.2.1 Different hypotheses for pancrustacean phylogeny

In recent years, numerous molecular phylogenetic analyses have been published, with a broad enough sample of crustacean taxa to allow a detailed placement of the insects within the Pancrustacea. Of particular importance have been analyses based on three different datasets: 1. the three nuclear protein coding genes elongation factor-1α (EF-1α), RNA polymerase II (PolII) and elongation factor-2 (EF-2) (Regier and Shultz, 2001, Regier, *et al.*, 2005), 2. complete nuclear ribosomal RNAs 18S and 28S (Mallatt and Giribet, 2006, Mallatt, *et al.*, 2004), and 3. the mitochondrial protein coding genes (for example Carapelli, *et al.*, 2007, Cook, *et al.*, 2005, Hassanin, 2006, Hassanin, *et al.*, 2005, Lavrov, *et al.*, 2004, Nardi, *et al.*, 2003). However, rather than clarifying the position of the insects, there has been a lack of consensus between these analyses. For a summary see figure 3.1.

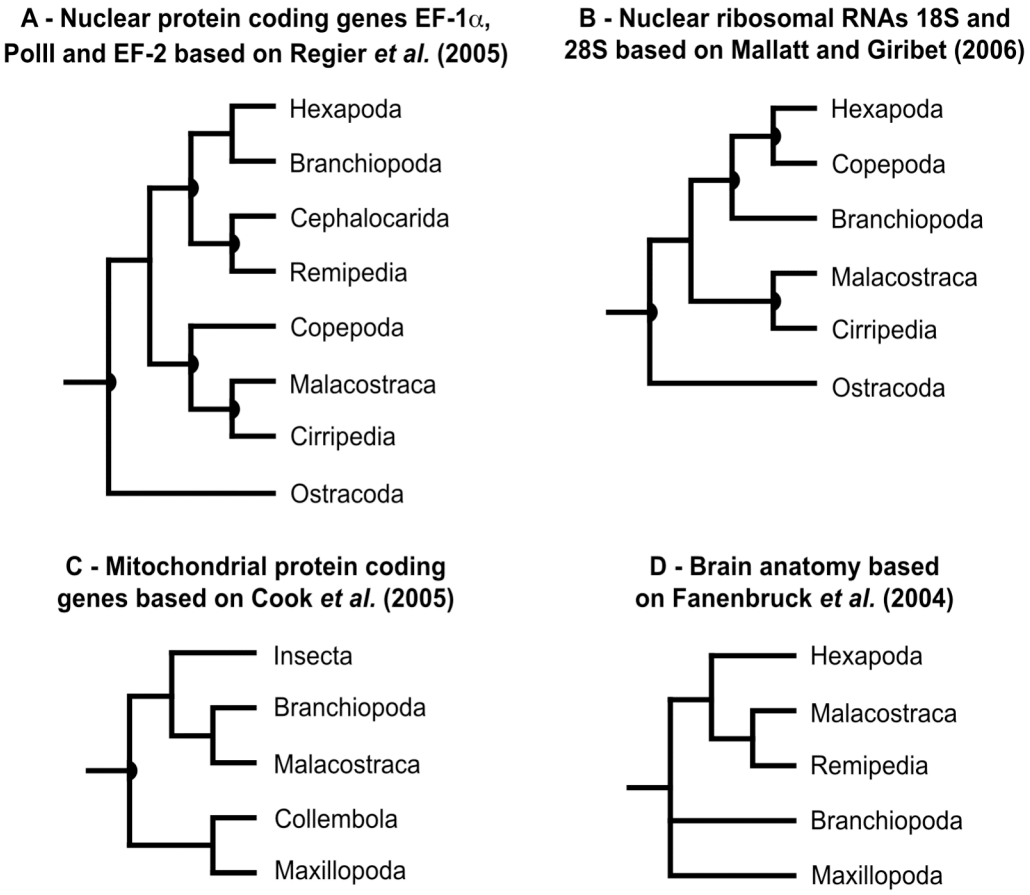


Figure 3.1. Hypotheses for the phylogeny of the Pancrustacea favoured by analyses of different datasets. Schematics showing the relationships between the major pancrustacean taxa favoured by (A) nuclear protein coding genes (EF-1α, Pol II and EF-2) as seen in the analyses of Regier *et al*. (2005) (based on the maximum likelihood analysis of amino acid sequences shown in their figure 1) and (B) nuclear ribosomal RNAs (complete 18S and 28S) as seen in the analyses of Mallatt and Giribet (2006) (based on the maximum likelihood analysis shown in their figure 1). The similar topologies supported by these analyses are not recovered by the analyses of (C) mitochondrial protein coding genes, as seen in Cook *et al*. (2005) (based on the Bayesian analysis of an amino acid dataset shown in their figure 4). Pancrustacean phylogeny has also been inferred from (D) neurobiology largely based on brain anatomy, as seen in Fanenbruck *et al*. (2004) (based on their figure 4). Where applicable (A, B, C), nodes that the authors claim to be supported by non-parametric bootstrapping are indicated with black half circles. For a more detailed treatment of the phylogenetic hypotheses for the Pancrustacea recovered by different datasets and the support for the different groupings, see section 3.2.1.

The most broadly sampled analyses based on the nuclear rRNAs (Mallatt and Giribet, 2006) and the nuclear protein coding genes (Regier, *et al*., 2005) find largely the same topology for the Pancrustacea (see figure 3.1 A and B). The Bayesian and likelihood analyses of combined 18S and 28S rRNAs and Bayesian, likelihood and parsimony analyses of concatenated sequences for the three nuclear protein coding genes support grouping the insects with the other hexapod taxa (the collembolans and diplurans).

They support a closer relationship for these hexapods to the branchiopods than to the malacostracans, which themselves group with the cirripedes. Also, both analyses place the ostracods at the base of the Pancrustacea, although with weak support.

There are some differences in particular relating to the position of the copepods relative to the hexapods. The different methods of analysis of the nuclear protein coding genes all support grouping the copepods with the malacostracans and cirripedes, while the hexapods form an unresolved group with the branchiopods and the two enigmatic taxa: the remipedes and cephalocarids (these two taxa were not represented in the rRNA analyses). In contrast, the Bayesian and likelihood analyses of the rRNAs support a placement of the copepods as the sister taxon to the hexapods. However, Mallatt and Giribet (2006) question this position for the copepods. Their parametric bootstrap analyses could not reject a hexapod-branchiopod sister-grouping and these tests grouped the copepods with the malacostracans and cirripedes in the best alternative tree.

The topologies supported by these nuclear gene analyses are not recovered by the analyses based on mitochondrial genes (as illustrated in figure 3.1 C). Bayesian and likelihood analyses of nucleotide and amino acid sequences of mitochondrial protein coding genes (including analyses with a model of amino acid substitution based on pancrustacean mitochondrial genomes (Carapelli, *et al.*, 2007)) do not recover a monophyletic hexapod group; the insects group with the malacostracans and branchiopods, to the exclusion of the collembolans (Carapelli, *et al.*, 2007, Cook, *et al.*, 2005, Lavrov, *et al.*, 2004, Nardi, *et al.*, 2003). The insects tend to be recovered as the sister-group to a malacostracan-branchiopod clade, although some analyses place them as the sister-group to the malacostracans (Carapelli, *et al.*, 2007, Nardi, *et al.*, 2003). The maxillopod crustaceans are often recovered as a clade, sometimes as the sister-group to the collembolans (Cook, *et al.*, 2005, Lavrov, *et al.*, 2004), whilst the remipede and cephalocarid are generally unstable leading to their exclusion from some analyses (Cook, *et al.*, 2005). However, whilst the relationships supported by mitochondrial genes tend to receive strong support from Bayesian posterior probabilities, they receive poor bootstrap support and Cook *et al.* (2005) were unable to reject alternative hypotheses, most notably for hexapod monophyly.

Phylogenies based on mitochondrial genes have often recovered unexpected results and their use in phylogenetics has been questioned (Curole and Kocher, 1999). Recently, attempts have been made to address some of the biases in mitochondrial genomes which could be problematic for phylogenetic reconstruction. In particular, Hassanin *et al.* (2005) demonstrate that mitochondrial genomes have a strand asymmetry in their nucleotide composition (with one strand bias towards A and C, the other towards T and G), leading to an asymmetric mutational constraint. When either individual genes or the control region are reversed in their orientation, the mutational constraint reverses, changing the frequency of different mutation types within a gene. This leads to long-branch attraction artifacts between the taxa with the reversed mutational constraints. Hassanin *et al.* (2005) show that the removal of taxa with reversed mutational constraints, or recoding the neutral or quasineutral mutations (their "Neutral Transitions Excluded" model) addresses these long-branch attraction artefacts. They also recover a monophyletic hexapod group in some of their analyses. However, this grouping is not recovered in the taxonomically broader analysis of Hassanin (2006) which also addresses the problems of strand asymmetry, and there is no specific support for the close relationship between the hexapods and branchiopods which is supported by the nuclear analyses. The relationships between the different pancrustacean groups tend to resemble the previous mitochondrial analyses, again with low bootstrap support.

Nuclear genes have not been subjected to the criticism that mitochondrial genes have been. However, it has been difficult to find any convincing support from other sources for the close relationship between the hexapods and branchiopods supported by the two nuclear datasets. The other major source of evidence for pancrustacean phylogeny has been from neurobiology, and in particular from brain anatomy and the structure of the optic lobes. Phylogenetic reconstructions based on these data have supported neither the nuclear nor the mitochondrial based phylogenies (figure 3.1 D), instead supporting a grouping of the hexapods with the malacostracans and the remipedes, to the exclusion of branchiopods and maxillopod taxa – collectively referred to as the "Entomostraca" (Fanenbruck*, et al.*, 2004, Harzsch, 2002, Sinakevitch*, et al.*, 2003). It is, therefore, currently uncertain where the insects fall within the Pancrustacea.

3.2.2 Different approaches to multigene analysis

Recently, many phylogenetic problems such have been tackled by combining the data from the various different analyses into a single large analysis. However, there has been a degree of controversy as to how best to combine the different datasets. Two main approaches have been advocated: "supermatrix" and "supertree" (Bininda-Emonds, *et al*., 2003, Gatesy, *et al*., 2002). Under the supermatrix approach the primary source data (for example sequences or morphological characters) are combined into a single matrix and analysed simultaneously. In contrast, under the supertree approach the topologies supported by different datasets are encoded into a matrix and used to generate a tree. There has been a degree of controversy as to which mode of analysis is more appropriate.

One of the main criticisms the supertree approach has received is that primary characters are duplicated amongst the source trees (Gatesy, *et al*., 2002). Also, there has been criticism for the inclusion of poor quality source trees, for example by poorly justified trees or trees with *a priori* phylogenetic constraints. It has been argued that these factors result in trees which can be "imprecise summaries of previous work" (Gatesy, *et al*., 2002). In contrast, the supermatrix approach has been criticised for discarding useful sources of phylogenetic data (Bininda-Emonds, *et al*., 2003). Phylogenetic hypotheses which are not based on character data cannot be coded into a matrix and their emission represents a loss of phylogenetic information. Also, supermatrices can be computationally more complex to analyse than supertrees, especially when they include different models for data from different sources (see section 3.2.3).

Bininda-Emonds *et al*. (2003) argue that both supermatrix and supertree approaches are useful summaries of their respective source data. Depending on the choice of tree weighting and the inclusion of source trees, supertrees can be precise summaries of previous work. On the other hand, the computational complexity associated with the supermatrix approach is being reduced. The advent of Bayesian phylogenetics has made the analysis of large datasets more tractable (Huelsenbeck, *et al*., 2001). I take a

supermatrix approach to address the question of pancrustacean phylogeny and the position of the insects.

I have compiled the different datasets previously used to address arthropod phylogeny to produce the largest multigene supermatrix dataset yet used for a Bayesian analysis of the Pancrustacea. Although all the major crustacean and hexapod groups are represented although not every gene is represented for every taxon, Philippe *et al.* (2004) have shown that even an important amount of missing data (for example 25%) is only a minor problem for likelihood analyses of large datasets. Moreover, Wiens and Moen (2008) demonstrate that in a Bayesian framework taxa with up to 95% missing data can be accurately placed provided the overall number of characters is large. However, whilst phylogenetic analyses can cope with missing data, where possible I have added new sequences to fill important gaps in the previous datasets. However, as I will now discuss, there are a number of important factors to consider when analysing such multigene datasets.

3.2.3 Considerations when analysing a multigene dataset

Multigene datasets often combine genes from a range of different sources and these are likely to evolve under different pressures and constraints (Castoe, *et al.*, 2004). Consider the different sources of data used to analyse arthropod phylogeny: mitochondrial genes appear to evolve at much higher rates (Curole and Kocher, 1999) and have different nucleotide compositions to nuclear genes which can affect phylogenetic analyses (Hassanin, *et al.*, 2005). Similarly, protein coding genes appear to evolve under different constraints compared to rRNAs. For protein coding genes, substitutions are under different constraints at different codon positions (Bofkin and Goldman, 2007). For rRNAs, substitutions are constrained by base pairing in paired "stem" regions but not in unpaired "loop" regions (Telford, *et al.*, 2005). Therefore, it is unlikely that one model can account for the heterogeneities in such a multigene dataset.

A common way of dealing with these heterogeneities has been to group sites evolving under similar pressures into predefined partitions (Castoe, *et al.*, 2004). The different partitions are free to evolve under different models. For some types of data there is strong support for particular partitioning strategies and models of analysis. For example, when analysing rRNA sequences there is evidence that sites belonging to stem and loop regions should be placed into separate partitions (Telford, *et al.*, 2005), and stem sites should be modelled with a doublet model to account for the constraints of base pairing (Schöniger and von Haeseler, 1994).

However, the most appropriate partitioning strategy is not always clear. For example, there is little consensus as to how to model the heterogeneities between codon positions. Some authors advocate removing the third codon position as the high rate of substitution at this position means that it is often saturated whilst others advocate placing the different codon positions in different partitions (for example Brandley, *et al.*, 2005, Regier and Shultz, 2001). It is also unclear how best to allocate models of nucleotide substitution to different partitions (Nylander, *et al.*, 2004). Incorrect modelling of a dataset can lead to problems associated with both overly simple models (underparameterisation) and overly complex models (overparameterisation) (Lemmon and Moriarty, 2004).

Fortunately, a number of different decision criteria can be used to select the most appropriate modelling strategy for a particular dataset, such as the Bayes factors and the Akaike information criterion (Posada, 2003, Posada and Buckley, 2004). Also, the phylogenetic signal in a set of aligned sequences can be analysed, which can indicate whether parts of a dataset, such as the third codon position, should be removed: likelihood-mapping (Strimmer and vonHaeseler, 1997) or saturation plots (as in Negrisolo, *et al.*, 2004) can be used to investigate the level of phylogenetic signal within a dataset, and factors such as compositional heterogeneity, which may bias phylogenetic reconstruction (Hassanin, *et al.*, 2005) can be examined.

3.2.4 Problems with convergence

There are also potential problems relating to running Bayesian analyses on large (multitaxon) datasets. One particular problem relates to the convergence of the Markov chain Monte Carlo (MCMC) on the posterior distribution. In theory, an MCMC will eventually converge on the posterior probability distribution, but in practice, there can be various difficulties (Beiko, *et al*., 2006, Huelsenbeck, *et al*., 2002). For example, an MCMC can get stuck in one region of parameter space before reaching the posterior distribution; the chain oscillates around what appears to be a stable likelihood value before starting to climb to higher likelihood values. Also, there can be problems with the MCMC mixing through the posterior probability distribution. The chain can get trapped on a single mode of a multimodal distribution and therefore not sample the entire distribution. For datasets with over 30 taxa, multimodality has been shown to be a problem (Beiko, *et al*., 2006). There are various criteria for judging when the MCMC has converged on the posterior distribution, such as graphically inspecting log likelihood values or comparing the tree topologies sampled from the posterior distribution. However, it is often said that whilst convergence is easy to reject it is "impossible" to accept (Beiko, *et al*., 2006).

I present an analysis of pancrustacean phylogeny, taking these various considerations into account. I run a number of analyses examining which is the most suitable way to model the data. First I investigate the effect of different treatments of the codon positions. I analyse the signal at each position to assess whether there is any *a priori* basis for favouring a particular treatment and I run a number of analyses under different partitioning strategies, using a number of decision criteria to select the most suitable model for the data. I then investigate the effect of analysing the data under different models of nucleotide substitution, again using different decision criteria to select the most appropriate modelling strategies. I also investigate the effect of analysing the data coded as an amino acid sequence. For all analyses, I examine whether the MCMC has converged on the posterior distribution. I then specifically address the question of the position of the insects within the Pancrustacea. Using the most suitable modelling strategy for the data, I use Bayes factors to carry out a set of hypothesis tests. I investigate whether the grouping of the hexapods with the branchiopod crustaceans that

is supported by the different analyses I have run, is favoured over alternative placements of the hexapods in the Pancrustacea.

3.3 Materials and Methods

3.3.1 Compiling a multigene dataset for analysing pancrustacean phylogeny

A multigene dataset for analysing pancrustacean phylogeny was compiled as described in section 2.2.1. Arthropod sequences for a range of genes were downloaded from GenBank and organised into monophyletic groups representing a diversity of crustacean taxa (section 2.2.2). Genes were sequenced to fill in gaps in the dataset (section 2.2.3) and the sequences of the different genes were concatenated into multigene sequences representing the different crustacean groups (section 2.2.4).

3.3.2 Gene by gene analysis of the dataset

The dataset was analysed to estimate the proportion of invariant sites and the nucleotide frequency for each gene. In PAUP* (Swofford, 2002), a neighbour joining tree was generated for each gene using a GTR model for the distance option with default settings. Using each tree, likelihood scores were then evaluated for that gene. The likelihood settings for the analyses were set to estimate the rate matrix for a GTR, the nucleotide frequencies, the proportion of invariant sites and the shape parameter for a four category gamma distribution.

3.3.3 Analysis of the phylogenetic signal at the different codon positions

The phylogenetic signal at the three different codon positions of the nuclear and mitochondrial partitions was analysed using likelihood-mapping, saturation plots and nucleotide composition plots as described in section 2.2.5.

3.3.4 Phylogenetic analysis and convergence on the posterior distribution

Bayesian phylogenetic analyses were run as described in section 2.2.6. The criteria used for terminating the analysis do not guarantee that the MCMC will have converged on the posterior distribution. To confirm that the topologies sampled by the MCMC were repeatable a second independent run was carried out for each analysis. The second MCMC was run for the same number of generations as the first, and for longer if it did not appear to have reached a plateau in that time. The extent of topological convergence between the two runs was assessed by calculating the split frequencies (see section 2.2.7).

It is important to point out that non-convergence on a topology could result from one MCMC getting stuck at a lower likelihood distribution. Here the run with the higher likelihood could be an accurate reflection of the posterior distribution. However, caution should be exercised when interpreting these cases; in the absence of topological convergence, the topology seen in the higher likelihood run cannot be guaranteed to be representative of the posterior distribution. Non-convergence could also result from incomplete sampling of the posterior distribution by one or both MCMCs. In this situation, both runs would have reached a plateau at the same distribution of log likelihoods, but it is possible that neither run is an accurate reflection of the posterior distribution. To distinguish between these two situations, the distributions of log likelihoods were compared graphically (section 2.2.7).

As a rule, discussions of topology and comparisons between models are restricted to the run with the higher log harmonic mean of the likelihoods (output by the MrBayes *sump*

command) even if both runs showed good topological convergence and sampled similar log likelihood distributions.

3.3.5 Selecting the most appropriate model

Bayes factors were used to select the most appropriate model for the data (section 2.2.8). Several studies have found that Bayes factors support parameter rich models. However, it is not clear whether this is truly because the data are very heterogeneous and the additional parameters are required to model it adequately, or whether Bayes factors tend to support the addition of parameters even when it is not necessary (Brown and Lemmon, 2007). Therefore, two additional selection criteria were used: the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). These criteria penalise the addition of parameters to a model. The AIC and BIC should be applied in a likelihood framework; therefore, estimates of the criteria were used derived from the Bayesian analyses (see section 2.2.9).

3.3.6 Tests of phylogenetic hypotheses

Bayes factors were used to carry out phylogenetic hypothesis tests. Bayes factors were calculated as in section 2.2.8; the different models were Bayesian analyses, constrained to different tree topologies. In certain cases (see section 3.4.6) the marginal likelihood was also calculated by an alternative method, known as smoothing (Suchard, *et al.*, 2005). This is implemented in Tracer v1.4 (Rambaut and Drummond, 2007). The Bayes factor was calculated as in section 2.2.8, with the harmonic mean replaced by the smoothed marginal likelihood estimate.

For some applications (see section 3.4.6) it was also useful to know how variable the marginal likelihood estimate was, as this would affect the size of the Bayes factor. 95% confidence intervals were calculated for the marginal likelihood estimates. The marginal likelihood is estimated in Tracer v1.4 (either as a harmonic mean or a smoothed estimate). This calculates the standard error from a bootstrap analysis (using

1000 replicates). The 95% confidence interval was calculated as $l \pm 1.96SE$ where l is the marginal likelihood estimate and SE is the standard error.

3.4 Results

3.4.1 The dataset

To address pancrustacean phylogeny I compiled a dataset representing 41 taxa from across the arthropods. The dataset consists of 16370 nucleotide sites from three different sources: rRNAs (18S and 28S), nuclear protein coding genes (EF-1α, PolII, EF-2 and histone H3) and mitochondrial protein coding genes. As there is evidence that these different types of data evolve under different constraints (see section 3.2.3) they should be treated as different partitions in a phylogenetic analysis. Inspecting the nucleotide frequencies of the different genes by eye (table 3.1) supports a difference in evolutionary process between the different sources; the nuclear genes generally have all nucleotides at a frequency of about 25%, whilst mitochondrial genes have elevated levels of A and T. Also, the mitochondrial protein coding genes tend to have a lower proportion of invariant sites than the nuclear protein coding genes, suggesting a faster rate of substitution (the very low proportion of invariant sites for NAD2 appears to be the result of a divergent sequence for *Lepeophtheirus*, as indicated in table 3.1). The proportion of invariant sites for 18S is very low compared to the other nuclear genes (10.30% compared to over 30%). However, this seems to be largely due to a divergent sequence for *Speleonectes*. This taxon proved very difficult to align with the other 18S sequences, and when it is removed the estimated proportion of invariant sites increases to 20.14% (see table 3.1).

Importantly, all the major crustacean groups, and a range of hexapod taxa, including the entognathous hexapods are represented in the dataset. There are newly sequenced 28S rRNAs for four previously underrepresented groups: the Collembola

Table 3.1. Core data for the different genes comprising the multigene dataset. The gene length is given for each gene, as are the proportion of invariant sites and the frequency of each nucleotide as described in section 3.3.2. The other parameter values from these analyses (nucleotide substitution rates and gamma distribution shape paramter) are also given.

	Nuclear						Mitochondrial protein											
	rRNA		Protein															
	18S	28S	EF-1a	PolII	EF-2	H3	ATP6	COX1	COX2	COX3	CYTB	NAD1	NAD2	NAD3	NAD4	NAD4L	NAD5	NAD6
Gene length (nucleotide sites)	1510	2311	1131	2124	2064	324	345	1476	591	759	1065	618	282	189	711	45	765	60
Invariant sites (%)	10.30* (20.14)	34.77	35.25	35.38	39.22	50.34	9.24	29.00	10.69	17.02	21.08	20.33	1.00* (10.03)	22.32	13.72	18.44	22.91	15.05
Nucleotide frequency (%)																		
A	24.96	25.96	28.26	27.99	28.17	23.82	34.26	36.59	36.52	35.33	34.45	30.70	36.22	27.70	28.50	28.48	31.18	24.86
C	23.52	22.78	23.33	22.52	22.58	29.53	14.59	15.36	16.51	15.79	16.68	9.67	19.40	15.54	14.50	12.80	12.72	12.67
G	29.34	28.79	22.51	22.84	22.03	23.42	12.34	11.25	11.27	12.21	10.70	16.98	10.86	18.78	15.76	23.13	15.48	19.18
T	22.18	22.46	25.90	26.65	27.22	23.23	38.81	36.80	35.70	36.68	38.18	42.65	33.52	37.98	41.25	35.59	40.62	43.30
Substitution rate																		
A<->C	1.167	1.128	1.718	2.346	2.125	1.941	1.732	0.537	0.726	1.092	1.158	0.248	0.103	10.854	0.086	0.001	0.246	8.1E+4
A<->G	2.821	2.639	3.426	4.225	4.210	5.746	4.139	4.746	4.708	4.720	4.064	6.585	2.310	17.090	4.476	7.3E+5	4.893	1.2E+5
A<->T	1.097	1.217	2.160	2.064	2.329	4.518	0.906	0.628	0.888	1.510	0.552	1.112	0.380	9.885	1.183	2.9E+5	1.257	3.8E+4
C<->G	0.592	0.802	2.068	1.612	2.449	2.233	6.519	3.950	3.167	3.882	3.310	5.635	2.022	8.008	3.113	9.1E+5	4.203	4.7E+4
C<->T	4.787	5.227	6.148	7.739	7.814	7.330	10.950	7.035	6.151	9.357	7.705	6.164	1.406	44.816	2.851	8.0E+5	6.063	1.7E+5
G<->T	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Gamma distribution shape parameter	0.510	0.575	0.935	0.844	1.028	1.164	0.375	0.490	0.463	0.525	0.565	0.625	0.383	0.932	0.721	0.845	0.656	1.283

* For 18S and NAD2 the proportions of invariant sites were very low. Inspection of the alignments for these genes indicated that they included very divergent taxa (*Speleonectes* for 18S and *Lepeophtheirus* for NAD2). The proportions of invariant sites with these taxa removed are indicated in parentheses.

Table 3.2. Summary of taxa used in the multigene dataset and the make up of the sequences. Taxa are organised into the major pancrustacean groupings, and the length of sequence in total and in each of the different data types (nuclear rRNAs, nuclear protein encoding genes and mitochondrial protein encoding genes) is given. The total length of sequence for each type of data is also given, as is the average length of the sequence across all taxa. Complete or near complete sequences are shown in bold. Sequences including new data are in italics.

Taxa	Number of sites				
	Nuclear		Mitochondrial Protein	Total sequence	% complete
	rRNA	Protein			
Insecta					
Acrididae	2616	1456	**6893**	10965	67.0
Archaeognatha	2642	**5567**	**6906**	15115	92.3
Blattaria	**3820**	**5494**	**6892**	16206	99.0
Drosophila	**3782**	**5643**	**6906**	16331	99.8
Hexagenia	2551	**5306**	0	7857	48.0
Lepismatidae	**3821**	**5547**	**6906**	16274	99.4
Diplura					
Campodeoidea	2451	**5547**	0	7998	48.9
Japygoidea	1860	4637	**6892**	13389	81.8
Collembola					
Entomobryomorpha	***3669***	**5467**	0	9136	55.8
Podura	**3739**	4783	**6906**	15427	94.2
Branchiopoda					
Artemia	**3790**	**5129**	**6796**	15714	96.0
Daphniidae	**3772**	1111	**6906**	11789	72.0
Limnadiidae	2626	**5496**	0	8122	49.6
Triops	**3817**	**5528**	**6906**	16251	99.3
Malacostraca					
Leptostraca	2408	**5487**	0	7895	48.2
Oniscidea	***3707***	**5279**	0	8986	54.9
Reptantia	**3815**	**5508**	**6871**	16194	98.9
Stomatopoda	**3820**	**5509**	**6885**	16214	99.0
Cirripedia					
Balanidae	***3480***	**5419**	**6906**	15805	96.5
Lepas	***3634***	**5231**	0	8865	54.2
Pollicipes	2395	322	**6906**	9622	58.8
Sacculinidae	2476	**5377**	0	7853	48.0
Copepoda					
Calanoida	***3714***	**5220**	0	8934	54.6
Cyclopidae	**3789**	**5222**	0	9011	55.0
Lepeophtheirus	1506	0	**6634**	8140	49.7
Tigriopus	**3475**	322	5718	9515	58.1
Branchiura					
Argulus	**3740**	**5231**	**6878**	15849	96.8
Podocopa ("Ostracoda")					
Cypridiidae	**3778**	**5189**	0	8967	54.8
Myodocopa ("Ostracoda")					
Cypridinidae	2982	**5231**	**6630**	14843	90.7
Cephalocarida					
Hutchinsoniella	1673	**5305**	**6878**	13856	84.6

(Table 3.2 continued)

Taxa	Number of sites			Total sequence	% complete
	Nuclear		Mitochondrial Protein		
	rRNA	Protein			
Remipedia					
Speleonectes	1577	4643	**6871**	13092	80.0
Outgroup: Myriapoda					
Lithobius	**3513**	**5546**	**6871**	15931	97.3
Pauropodidae	1720	**5547**	0	7267	44.4
Scutigerellidae	**3358**	**5128**	0	8486	51.8
Spirostreptida	**3774**	**5212**	**6906**	15892	97.1
Outgroup: Chelicerata					
Limulus	**3820**	**5323**	**6885**	16029	97.9
Mastigoproctus	1714	**5206**	3453	10373	63.4
Mygalomorphae	**3819**	2663	**6827**	13310	81.3
Phalangida	1730	**5509**	0	7239	44.2
Pyncnogonida	**3802**	**5547**	3416	12765	78.0
Scorpiones	**3772**	2882	**6603**	13257	81.0
Total number of sites	3821	5643	6906	16370	100.0
Mean number of sites per taxon	3121	4629	4318	12067	73.7
% complete	81.7	82.0	62.5	73.7	

(Entomobryomorpha), the Cirripeida (Balanidae and *Lepas*), the Copepoda (Calanoida) and the Malacostraca (Oniscidae). There are still gaps in the dataset, with not every taxon having sequence data for every gene. However, the dataset was constructed so that across all the broadly recognised higher groupings (such as the major crustacean subgroups), every gene was present in at least one taxon. This is summarised in table 3.2.

3.4.2 Signal at different codon positions

Before analysing the data under different treatments of codon position, the signal at each codon position was investigated to see whether there was any *a priori* reason to favour a particular treatment for the different codon positions.

Phylogenetic content of signal

Likelihood-mapping showed that for the three codon positions in each of the nuclear and mitochondrial datasets there were comparable levels of signal (figure 3.2). All

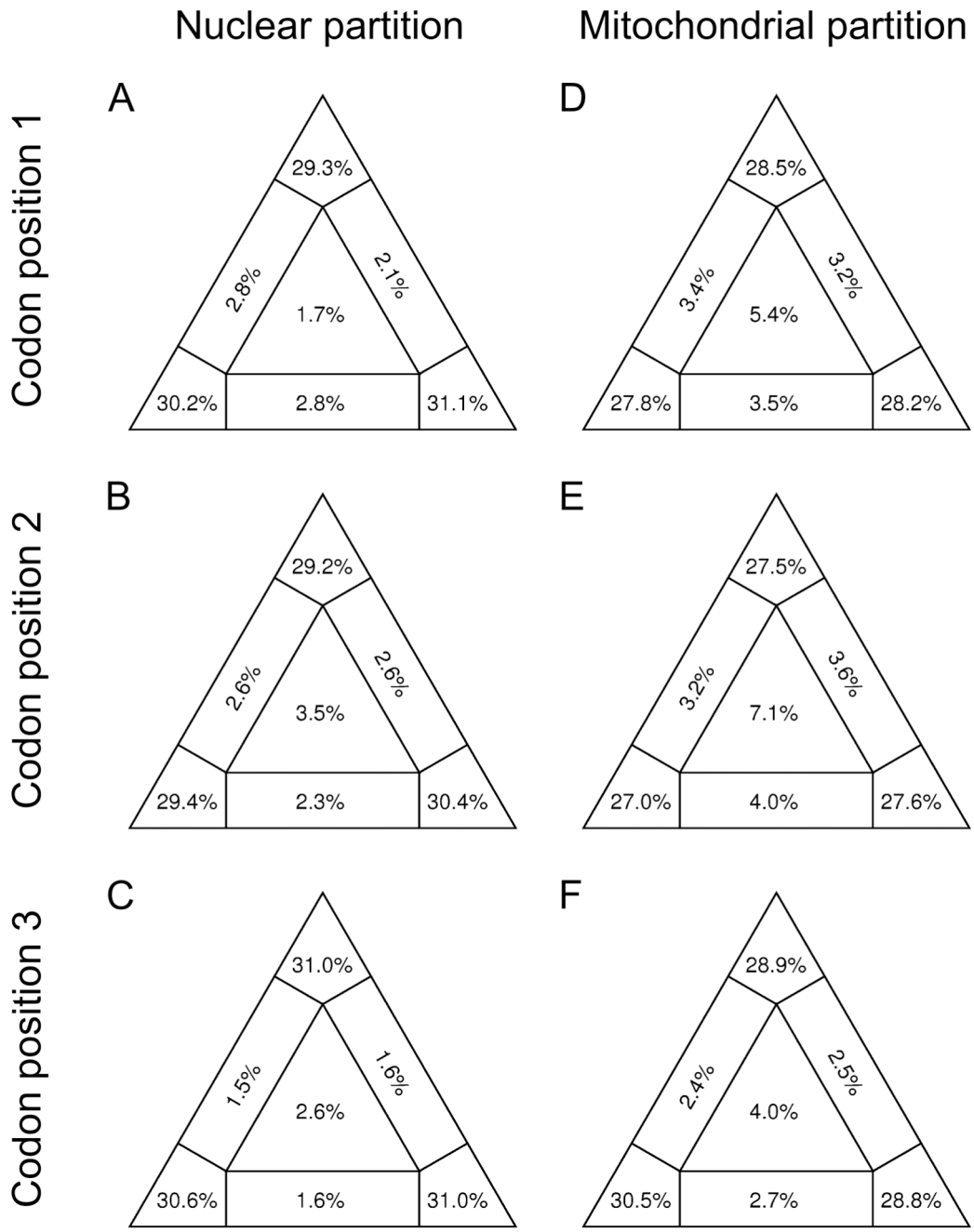


Figure 3.2. Likelihood mapping plots for each codon position of the nuclear and mitochondrial partitions. (A-C) nuclear partitions, (D-F) mitochondrial partitions, (A, D) first codon position, (B, E) second codon position, (C, F) third codon position. The percentages give the proportion of points falling in the different regions. For both the nuclear and mitochondrial partitions all codon positions show a majority of fully resolved quartets (points falling in the corners of the plots) and there is no obvious difference in signal between the different codon positions. For the nuclear partition all three codon positions (A-C) have approximately 90% fully resolved quartets and for the mitochondrial partition all three codon positions (D-F) have 80-90% fully resolved quartets.

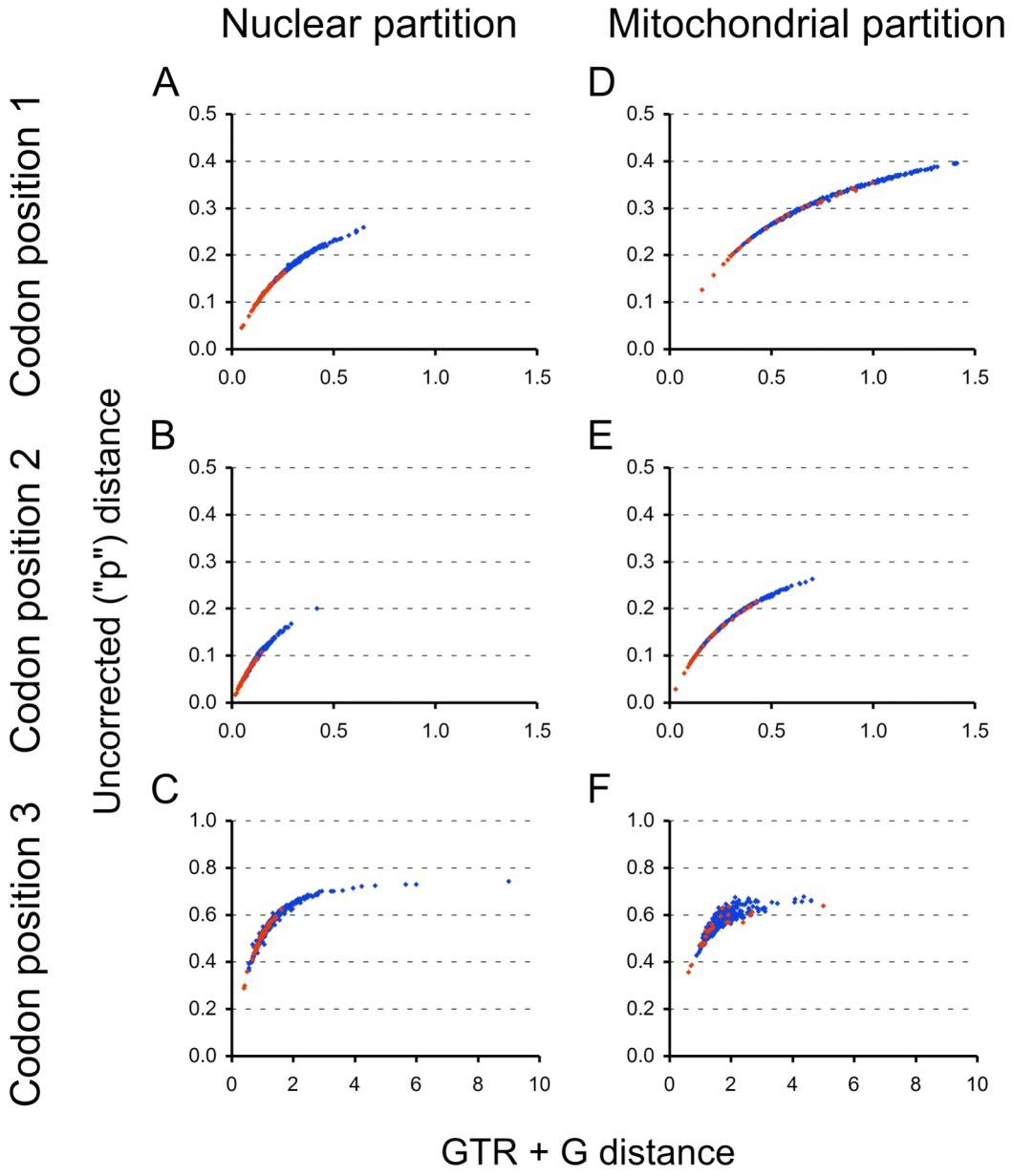


Figure 3.3. Saturation plots for each codon position of the nuclear and mitochondrial partitions. (A-C) nuclear partitions, (D-F) mitochondrial partitions, (A, D) first codon position, (B, E) second codon position, (C, F) third codon position. Taxon pairs belonging to morphologically well-supported groups (Insecta, Diplura, Collembola, Branchiopoda, Malacostraca, Cirripedia, Copepoda, Myriapoda and Chelicerata) are shown in red. All plots show a linear relationship between the GTR + G distance and the uncorrected ("p") distance, indicating the presence of signal in the data. For the two third codon positions (C, F), the plots level off at an uncorrected ("p") distance of 0.75 indicating a degree of signal saturation. Additionally, for both nuclear and mitochondrial partitions, the distance measures for the second codon positions (B, E) are shorter than for the first codon positions (A, D) suggesting heterogeneity in the evolutionary process between the codon positions.

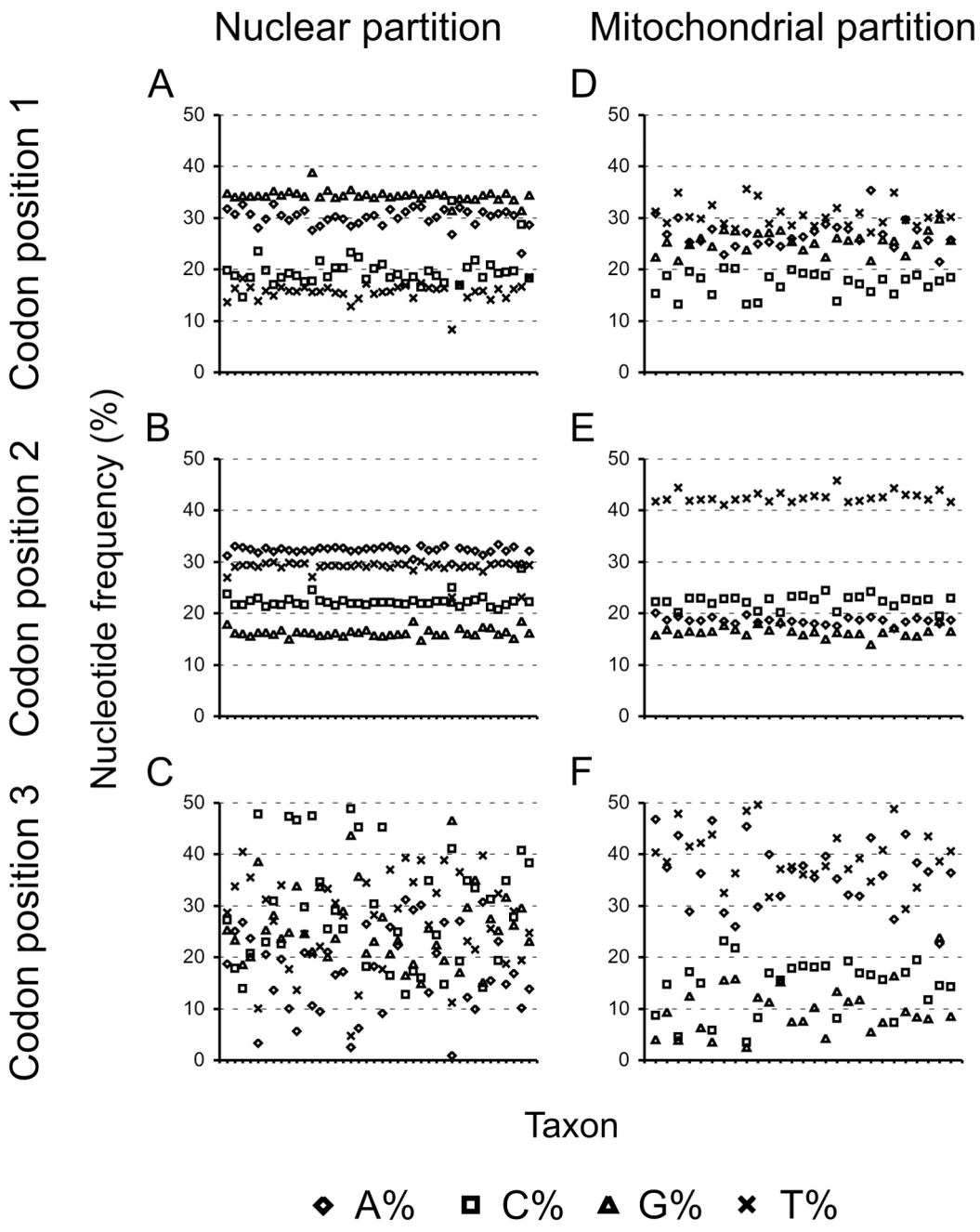


Figure 3.4. Composition plots for each codon position of the nuclear and mitochondrial partitions. (A-C) nuclear partitions, (D-F) mitochondrial partitions, (A, D) first codon position, (B, E) second codon position, (C, F) third codon position. For both nuclear and mitochondrial partitions, composition is most homogeneous at the second codon position (B, E) as seen in the relatively flat plots of nucleotide frequency across taxa. In contrast, composition is most heterogeneous at the third codon position (C, F) as seen in the large variability in the frequency of each nucleotide across taxa. Taxa are in alphabetical order. For the nuclear partition *Lepeophtheirus* was omitted, as there was no sequence data in the partition. For the mitochondrial partition, the following taxa were omitted as there was no sequence data in the partition: Calanoida, Campodeoidea, Cyclopidae, Cyprididae, Entomobryomorpha, *Hexagenia*, *Lepas*, Leptostraca, Limnadiidae, Oniscidea, Pauropodidae, Phalangida, Sacculinidae, Scutigerellidae.

three nuclear codon positions had over 90% fully resolved quartets, and all three mitochondrial codon positions had between 80% and 90% fully resolved quartets. This was largely mirrored in the saturation plots (figure 3.3). These plots show that for the first and second codon positions of both the nuclear and mitochondrial partitions, there was no saturation of signal. As expected, there appears to be some heterogeneity between these codon positions, as for both nuclear and mitochondrial genes, the distances are wider for the first codon positions than for the second codon positions. For the third codon positions, in both the nuclear and mitochondrial partitions, a large amount of the data fell on a slope indicating that there was signal in the data. However, there was a degree of saturation, as both plots levelled out at an uncorrected ("p") distance of 0.75. Therefore, whilst likelihood-mapping gives no basis for removing the third codon positions, saturation plots do give some support for their removal.

Nucleotide composition

Plots of nucleotide composition (figure 3.4) show that for both the nuclear and mitochondrial partitions, at the second codon positions nucleotide frequencies are largely homogeneous across taxa. For both nuclear and mitochondrial datasets, composition appears less conserved at the first codon positions, with the mitochondrial first position seemingly more heterogeneous than the nuclear. For both types of data, the third positions appear very compositionally heterogeneous. This supports the removal of the third codon positions, as compositional heterogeneity could potentially be problematic for phylogenetic reconstruction. These results also suggest that there are compositional differences between the codon positions.

3.4.3 Comparison of different modelling strategies

Comparison of the different codon partitioning strategies

The above analyses of signal at the different codon positions give some support to removing the third codon positions. However, as this was not unanimously supported by the different analyses, I ran Bayesian phylogenetic analyses under different

Table 3.3. Summary of the models used in the different analyses of pancrustacean phylogeny. The partitions used and the substitution models used in each partition are indicated.

Model name	rRNA		Nuclear protein encoding			Mitochondrial protein encoding		
	Stem	Loop	Codon position 1	Codon position 2	Codon position 3	Codon position 1	Codon position 2	Codon position 3
Comparisons of partitioning models								
Complete dataset								
C_1-GTR+G	GTR+G with doublet model	GTR+G	GTR+G			GTR+G		
C_2-GTR+G	GTR+G with doublet model	GTR+G	GTR+G		GTR+G	GTR+G		GTR+G
C_3-GTR+G	GTR+G with doublet model	GTR+G	GTR+G	GTR+G	GTR+G	GTR+G	GTR+G	GTR+G
Third codon position removed								
R_1-GTR+G	GTR+G with doublet model	GTR+G	GTR+G		-	GTR+G		-
R_2-GTR+G	GTR+G with doublet model	GTR+G	GTR+G	GTR+G	-	GTR+G	GTR+G	-
Comparisons of substitution models								
R_2-GTR	GTR with doublet model	GTR	GTR	GTR	-	GTR	GTR	-
R_2-HKY+G	HKY+G with doublet model	HKY+G	HKY+G	HKY+G	-	HKY+G	HKY+G	-
R_2-HKY	HKY with doublet model	HKY	HKY	HKY	-	HKY	HKY	-
Amino acid encoded proteins								
Amino acids	GTR+G with doublet model	GTR+G	Reversible jump model			Reversible jump model		

partitioning models of codon position with and without the third positions (table 3.3). The complete dataset (including the third positions) was analysed under three different partitioning models (C_1, C_2 and C_3 in table 3.4). In addition, a reduced dataset with the third positions excluded was analysed under two different partitioning models (R_1 and R_2 in table 3.4). Substitution rates in each partition were modelled using a GTR with a four category gamma distribution to avoid the risks of underparameterisation. A proportion of invariable sites was not included in the model as it has been argued that the gamma distribution is sufficiently general to allow for very low rates at some sites (Yang, 1996).

Table 3.4. Alternative partitioning strategies for dealing with heterogeneities between the different codon positions. Partitioning strategies are given for the complete dataset and a reduced dataset with the third codon position removed.

Partitioning name	Partitions	Number of partitions
Complete dataset		
C_1	rRNA stem, rRNA loop, nuclear protein all codon positions, mitochondrial protein all codon positions	4
C_2	rRNA stem, rRNA loop, nuclear protein first and second codon positions, nuclear protein third codon positions, mitochondrial protein first and second codon positions, mitochondrial protein third codon positions	6
C_3	rRNA stem, rRNA loop, nuclear protein first codon positions, nuclear protein second codon positions, nuclear protein third codon positions, mitochondrial protein first codon positions, mitochondrial protein second codon positions, mitochondrial protein third codon positions	8
Third codon position removed		
R_1	rRNA stem, rRNA loop, nuclear protein all codon positions, mitochondrial protein all codon positions	4
R_2	rRNA stem, rRNA loop, nuclear protein first codon positions, nuclear protein second codon positions, mitochondrial protein first codon positions, mitochondrial protein second codon positions	6

The fit of the different modelling strategies to the data for the complete and reduced datasets were compared using Bayes factors. Comparison of the three different partitioning strategies for the complete dataset showed that increasing the partitioning of the dataset increases the fit of the model to the data (see table 3.5). Increasing

Table 3.5. Bayes factors and estimates of AIC and BIC for comparisons between different modelling strategies. Model comparisons are given as $Model_2/Model_1$. Model comparisons are organised by whether they compared between different partitioning models of the codon positions or between different models of nucleotide substitution.

Models	$2\ln(BF_{21})$	Δ(AIC)	Δ(BIC)
Comparisons of partitioning models			
Complete dataset			
C_3-GTR+G / C_2-GTR+G	3064.51	3020.51	2851.04
C_3-GTR+G / C_1-GTR+G	10990.30	10902.30	10563.36
C_2-GTR+G / C_1-GTR+G	7925.79	7881.79	7712.32
Third codon position removed			
R_2-GTR+G / R_1-GTR+G	2748.25	2704.25	2541.28
Comparisons of substitution models			
R_2-GTR+G / R_2-HKY+G	1948.45	1888.45	1666.21
R_2-GTR+G / R_2-GTR	33877.97	33865.97	33821.52
R_2-GTR+G / R_2-HKY	40972.61	40900.61	40633.92
R_2-HKY+G / R_2-GTR	31929.52	31977.52	32155.31
R_2-HKY+G / R_2-HKY	39024.16	39012.16	38967.71
R_2-GTR / R_2-HKY	7094.64	7034.64	6812.40

partitioning also improved the fit of the model to the data for the reduced dataset. All Bayes factors were orders of magnitude greater than the cutoff value of 10. Estimates of the AIC and BIC also supported increasing the partitioning of the dataset (see table 3.5). As with the Bayes factors, all the differences in AIC and BIC were orders of magnitude above their cutoff values of 10. This supports the use of the fully partitioned models (C_3 and R_2).

As there were potential problems with the convergence of the MCMC on the posterior distribution, it was necessary to examine whether the two runs of each model had converged on the posterior distribution. For both models of the reduced dataset (R_1 and R_2) the two runs appear to have converged on the posterior distribution. For each model both runs sampled similar distributions of log likelihoods (figure 3.5) and the sampled topologies were similar as indicated by the split frequencies (table 3.6). This suggests that the topologies of the consensus trees are accurate reflections of the posterior distributions.

In contrast for all three models of the complete dataset (C_1, C_2 and C_3), the alternative runs appear to have reached plateaus at lower distributions of log likelihoods, notably so

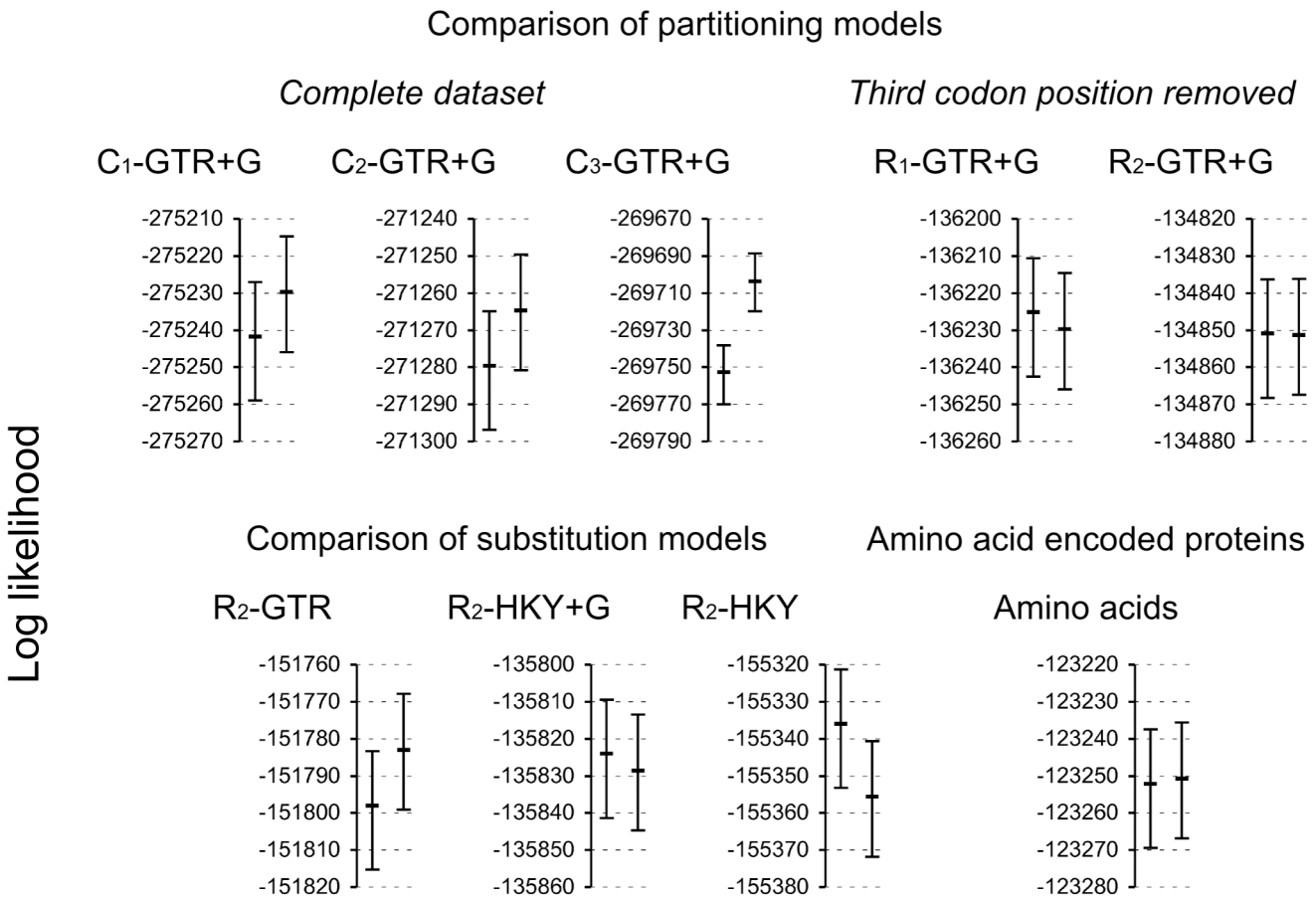


Figure 3.5. Comparisons of the distributions of log likelihoods for the two runs of each modelling strategy. For each modelling strategy the arithmetic mean of the post burnin log likelihood values is plotted for the two runs. Bars show the range of the middle 95% of the log likelihood values. The different models are summarised in table 3.2. The extent of the overlap of the two distributions gives an indication of whether they have sampled the same distribution of log likelihood values. For the *R_1-GTR+G*, *R_2-GTR+G*, *R_2-HKY+G* and *Amino acids* models, the two runs appear to have sampled similar distributions of log likelihoods. For the other models, one run appears to have sampled a lower distribution of log likelihoods than the other.

Table 3.6. Split frequencies for the different modelling strategies. Elevated values (>0.05) suggesting notable differences in the topologies sample are shown in bold. Models are organised by whether they addressed different partitioning models of the codon positions, different models of nucleotide substitution or coding the protein sequences as amino acids.

Model name	Split frequency
Comparisons of partitioning models	
Complete dataset	
C_1-GTR+G	0.005
C_2-GTR+G	0.040
C_3-GTR+G	**0.160**
Third codon position removed	
R_1-GTR+G	0.006
R_2-GTR+G	0.020
Comparisons of substitution models	
R_2-GTR	**0.108**
R_2-HKY+G	0.009
R_2-HKY	**0.222**
Amino acid encoded proteins	
Amino acids	**0.070**

for model C_3 (figure 3.5). This suggests that at least one of the runs did not converged on the posterior probability distribution. However, for the C_1 model the split frequency was low, indicating that the two runs still recovered a similar topology. This suggests that the recovered topology may still be an accurate reflection of the posterior distribution. In contrast, the C_2 and C_3 models had elevated split frequencies, especially for model C_3 (table 3.6). As the topologies of the preferred runs were not recovered by the alternative runs, they may not be reliable reflections of the posterior distributions.

The most appropriate treatment of the different codon positions therefore appears to be partitioning the codon positions, as for both the complete and reduced dataset the most partitioned models were supported by the different decision criteria used, and to remove the third codon position, as the runs with the third position included appeared to have problems converging on the posterior distribution.

Comparison of models of nucleotide substitution

All the above analyses were run modelling nucleotide substitutions in every partition using a GTR model with a gamma distribution to avoid the problems of underparameterisation. However, this is a highly parameter rich model which could potentially lead to problems resulting from overparameterisation (Lemmon and Moriarty, 2004). I therefore also ran Bayesian analyses modelling the nucleotide substitution rate in each partition with less parameter rich models, namely: GTR without a gamma distribution, HKY with a gamma distribution and HKY without a gamma distribution (table 3.7). All analyses were run using the reduced dataset, and the R_2 partitioning model (table 3.3), as this treatment was favoured by the above comparisons.

Table 3.7. Alternative models of nucleotide substitution and the number of model parameters. Number of parameters are given for the model alone, and for the whole modelling strategy when used with the R_2 partitioning strategy.

Substitution model	Number of parameters in substitution model	Total parameters under R_2 partitioning strategy
GTR+G	7	78
GTR	6	72
HKY+G	3	48
HKY	2	42

Bayes factors comparisons of the different models of nucleotide substitution showed that the most parameter rich model (GTR+G) was the most appropriate model for the dataset (table 3.5), receiving strong support over the next best model (HKY+G). However, the addition of parameters alone did not improve the fit of the model to the data, as modelling the data with an HKY+G (48 parameters) received strong support over the more parameter rich GTR model (72 parameters). Modelling the data with a GTR in turn received strong support over the HKY model. In all cases, Bayes factors were orders of magnitude greater than the cutoff value of 10. Estimates of the AIC and BIC gave the same relative support for the different models (see table 3.5). As with the Bayes factors, all the differences in AIC and BIC were orders of magnitude above their cutoff value of 10.

As with the GTR+G model, the two runs of the HKY+G model appear to have converged on the same posterior distribution. They sampled similar log likelihood values (figure 3.5), and they recovered very similar topologies, as indicated by low split frequencies (table 3.6). In contrast, for both the GTR and HKY models, the alternative run reached a plateau at a lower set of log likelihood values (figure 3.5) and the two runs sampled different sets of topologies indicated by the elevated split frequencies (table 3.6). None of the alternative models of nucleotide substitution appear preferable to the GTR+G.

Coding proteins as amino acid sequences

The dataset was also analysed with the protein coding genes recoded as amino acid sequences (table 3.3). The MCMC was set to select the most suitable model of amino acid substitution for the nuclear and mitochondrial partitions. Topological convergence appeared poor, reflected by elevated split frequencies (0.070; table 3.6). Surprisingly, the log likelihoods appeared to sample similar values (figure 3.5) suggesting that the two runs may have converged on the same posterior distribution but not sampled it adequately. Due to better convergence, the analysis of the favoured model coded as a nucleotide sequence appears more compelling than the analyses coded as an amino acid sequence.

3.4.4 Pancrustacean phylogeny

The phylogeny of the Pancrustacea supported by the favoured model of analysis (R_2-*GTR+G*) is shown if figure 3.6. Despite the differences in the fit of the different models to the data, a number of groupings were supported across all modelling strategies. These groups are summarised in table 3.8. All the analyses recovered the morphologically well-supported pancrustacean groups. The Diplura, Collembola, Branchiopoda, Malacostraca, Cirripedia and Copepoda were recovered with a Bayesian posterior probability (BPP) of 1.00, as were the outgroup taxa (Myriapoda and Chelicerata). The Pancrustacea were also recovered with a BPP of 1.00. The only anomalous result related to the Insecta. All models recovered the group with a BPP of

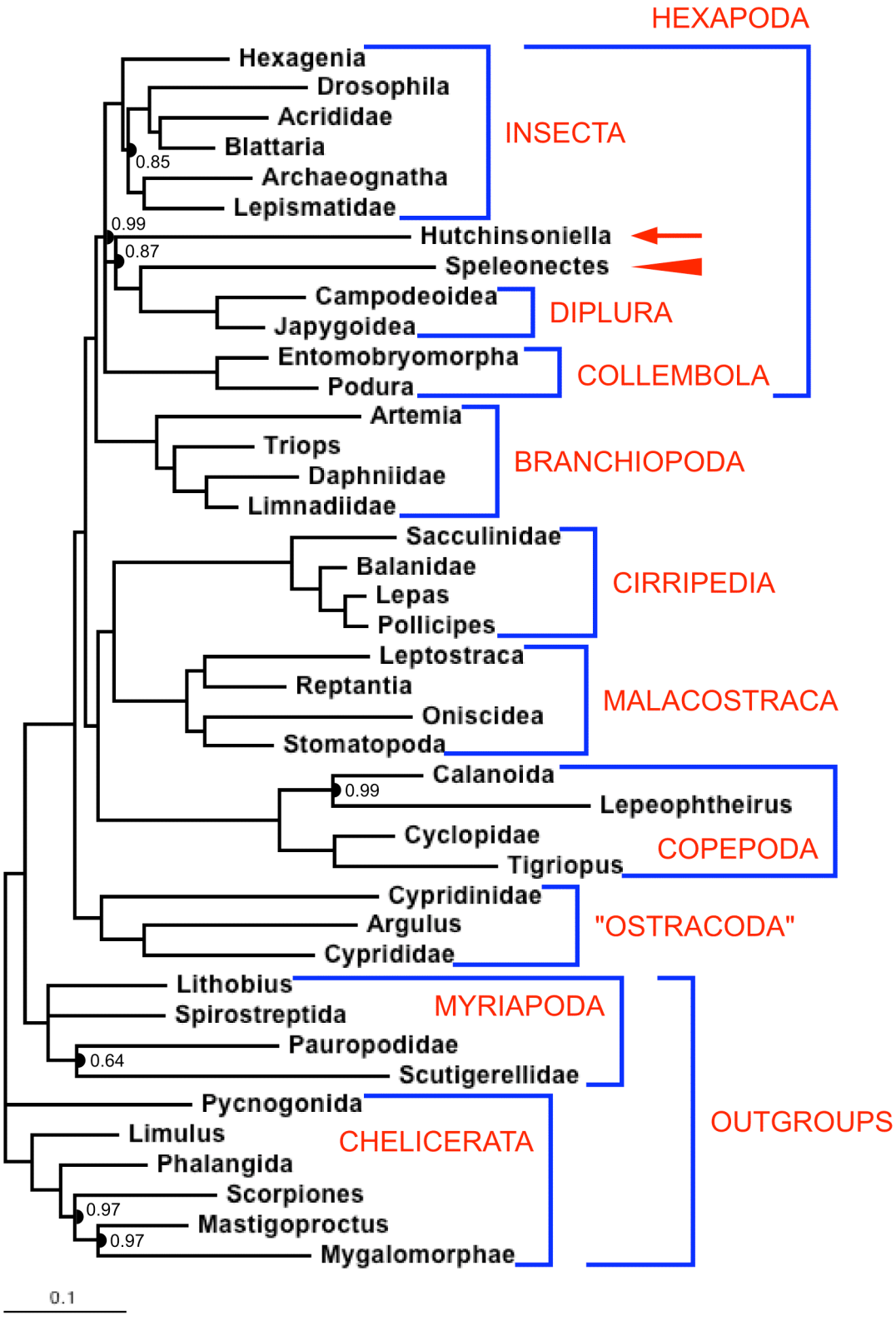


Figure 3.6. Consensus tree showing pancrustacean phylogeny analysed under the *R_2-GTR+G* model. The analysis used the reduced dataset (third codon position excluded) and a GTR+G model of nucleotide substitution; see table 3.3. The major pancrustacean groups are marked, as are the outgroups. The positions of the remipede *Speleonectes* and the cephalocarid *Hutchinsoniella* within the hexapods are indicated with an arrowhead and an arrow respectively. All nodes receive Bayesian posterior probability support values of 1.00, apart from the nodes marked with black half circles where the support value is indicated.

Table 3.8. Support for groupings within the Pancrustacea across the different modelling strategies. Bayesian posterior probability support values are given for the pancrustacean groupings that are well supported across the different modelling strategies. *Speleonectes* and *Hutchinsoniella* are omitted from these groupings as there are reasons to believe that their positions are artefactual (see section 3.5.1). The sister-groups to these two taxa are shown below.

	Complete dataset			Third codon position removed		Simpler models			Amino acids
	C_1-GTR+G	C_2-GTR+G	C_3-GTR+G	R_1-GTR+G	R_2-GTR+G	R_2-HKY+G	R_2-GTR	R_2-HKY	
Support for major pancrustacean groupings									
Insecta	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00
Diplura	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Collembola	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Hexapoda	1.00	0.70	1.00	1.00	0.99	1.00	1.00	-	0.62
Branchiopoda	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Malacostraca	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Cirripedia	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Copepoda	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Ostracoda (Cyprididae + Cypridinidae + *Argulus*)	-	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Hexapoda + Branchiopoda	0.62	0.60	1.00	1.00	1.00	1.00	1.00	-	0.58
Malacostraca + Cirripedia	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Malacostraca + Cirripedia + Copepoda	1.00	0.74	1.00	1.00	1.00	1.00	1.00	-	0.84
Pancrustacea – “Ostracoda"	-	1.00	1.00	1.00	1.00	1.00	1.00	-	1.00
Pancrustacea	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Myriapoda	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Chelicerata	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Positions of *Speleonectes* and *Hutchinsoniella*									
Speleonectes + Diplura	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
Hutchinsoniella + *Speleonectes* + Diplura	-	-	-	1.00	0.87	0.99	-	-	-
Speleonectes + *Hutchinsoniella*	-	-	-	-	-	-	-	-	0.62
Hutchinsoniella + *Drosophila*	-	-	1.00	-	-	-	-	0.98	-
Hutchinsoniella + Cyprididae	1.00	-	-	-	-	-	-	-	-
Hutchinsoniella + Copepoda	-	-	-	-	-	-	1.00	-	-
Hutchinsoniella + Hexapoda + Branchiopoda	-	0.60	-	-	-	-	-	-	-

1.00, apart from the C_3-*GTR+G* and R_2-*HKY* models where *Hutchinsoniella* fell within the insects as the sister to *Drosophila* (although in both cases, this grouping received strong support, with a BPPs of 1.00 and 0.98 respectively).

The position of the insects

Almost all the analyses recovered a hexapod assemblage where the Insecta, Diplura and Collembola grouped together to the exclusion of the various crustacean groups. Only the R_2-*HKY* model failed to recover this grouping (see table 3.8). However, as was seen in section 3.4.3 this model showed one of the poorest fits to the data and the convergence diagnostics suggested that it might not have converged on the posterior distribution. In fact, the alternative run under the model recovered many differences in topology from the preferred run (see table 3.9).

Other than the R_2-*HKY* analysis, the only exception to this hexapod monophyly is that the remipede *Speleonectes* was repeatedly recovered within this hexapod assemblage as the sister-group of the Diplura with strong support (table 3.8). Only in the analysis of the *Amino acids* model was this grouping not recovered. Additionally, in the favoured R_2-*GTR+G* analysis *Hutchinsoniella* also grouped with this assemblage. This was also seen in the well-supported alternative models for the reduced dataset (R_1-*GTR+G* and R_2-*HKY+G*). The two best fitting analyses of complete dataset (C_3-*GTR+G* and C_2-*GTR+G*) and the *Amino acids* analysis recovered alternative positions for *Hutchinsoniella* (and *Speleonectes* for the *Amino acids* analysis) (see table 3.8), but these runs were demonstrated in section 3.4.3 to show poor convergence. Inspection of the consensus tree topologies for the alternative runs shows that the topologies only differ from the preferred runs in the placement of *Hutchinsoniella* (and *Speleonectes* for the *Amino acids* analysis) (see table 3.9). This suggests that these modelling strategies had particular difficulty in placing these taxa, making their position in the preferred runs questionable. The positions of *Speleonectes* and *Hutchinsoniella* as successive sister-groups to the Diplura (recovered by the favoured R_2-*GTR+G* analysis) appear the best supported.

Table 3.9. Topological differences in the relative positions of the major pancrustacean groups between the preferred run and the alternative run for each model. Different groupings recovered by the two runs are shown and taxa whose positions varied between the two runs are in bold. Only differences in the overall topology are shown, not differences in Bayesian posterior probability support values for clades, as it is difficult to interpret differences in support values when various taxa differ in their placements between the runs. Models are organised by whether they addressed different partitioning models of the codon positions, different models of nucleotide substitution or coding the protein sequences as amino acids.

Model name	Preferred run	Alternative run
Comparisons of partitioning models		
Complete dataset		
C_1-GTR+G	-	-
C_2-GTR+G	***Hutchinsoniella*** + Hexapoda + Branchiopoda	***Hutchinsoniella*** + Copepoda
C_3-GTR+G	***Hutchinsoniella*** + *Drosophila*	***Hutchinsoniella*** + "Ostracoda"
Third codon position removed		
R_1-GTR+G	-	-
R_2-GTR+G	-	-
Comparisons of substitution models		
R_2-GTR	***Speleonectes*** + Diplura	***Speleonectes*** + "Ostracoda"
R_2-HKY+G	-	-
R_2-HKY	***Hutchinsoniella*** + Copepoda	***Hutchinsoniella*** + *Drosophila*
	Collembola + Branchiopoda	**Collembola** + Copepoda
	"Ostracoda" + (Copepoda + Collembola)	**"Ostracoda"** at base of Pancrustacea
	(Malacostraca + Cirripedia) at base of Pancrustacea	**(Malacostraca + Cirripedia)** + (Copepoda + *Hutchinsoniella*)
Amino acid encoded proteins		
Amino acids	***(Hutchinsoniella + Speleonectes)*** + (Hexapoda + Branchiopoda)	***Speleonectes*** + Diplura ***Hutchinsoniella*** + Copepoda
	Copepoda + (Malacostraca + Cirripedia)	(("Hexapoda" + *Speleonectes*) + Branchiopoda) + (**Copepoda** + *Hutchinsoniella*) + (Malacostraca + Cirripedia)

This hexapod group (with or without *Speleonectes* and/or *Hutchinsoniella*) was generally strongly supported (table 3.8), with only the analysis of the *C_2-GTR+G* and *Amino acids* models receiving less than a BPP of 0.99. The analyses also all recovered the branchiopods as the sister-group of this hexapod group (with or without *Speleonectes* and/or *Hutchinsoniella*). This grouping was strongly supported in the analyses of *R_2-GTR+G*, *R_1-GTR+G*, *R_2-HKY+G*, *R_2-GTR* and *C_3-GTR+G* models,

although support was reduced in the analyses of the C_2-*GTR+G*, C_1-*GTR+G* and *Amino acids* models (BPP approximately 0.60).

Other features of pancrustacean phylogeny

The malacostracans consistently grouped with the cirripedes (BPP of 1.00 across all analyses), and the copepods were recovered as the sister-group to this clade (apart from in the unfavoured R_2-*HKY* analysis). This grouping was generally well supported (table 3.8), although it received weaker support from the analysis of model C_2-*GTR+G* where *Hutchinsoniella* also fell in the group as the sister taxon to the copepods, and in the analysis of the *Amino acids* model. All analyses recovered an assemblage of the two ostracod taxa with the branchiuran *Argulus* at the base of the Pancrustacea, with strong support, with the exception of the poorly favoured R_2-*HKY* model and the C_1-*GTR+G* model. The C_1-*GTR+G* model recovered the myodocopan ostracod (taxon Cypridinidae) at the base of the Pancrustacea with *Argulus*, whilst the podocopan ostracod (taxon Cyprididae) grouped with *Hutchinsoniella* and the malcostracan-cirripede-copepod assemblage. However, as was seen in section 3.4.3, although the two runs under this modelling strategy seem to have reliably converged on the same topology (judging by the split frequencies), it showed the poorest fit to the data for the analyses of the complete dataset.

3.4.5 Hypothesis tests

The above analyses consistently recover a sister-group relationship between the hexapods and the branchiopods. To test how well this hypothesis was supported, Bayes factors were used to compared the hexapod-branchiopod grouping to other possible placements of the hexapods in the Pancrustacea. The different positions to which the hexapods were constrained are summarised in table 3.10. As well as grouping the hexapods with each of the major pancrustacean groupings (hypotheses H_1-H_4), the hexapods were also constrained to group with various assemblages of crustaceans that were repeatedly supported in the previous analyses (hypotheses H_5 and H_6). Also,

Table 3.10. Different hypotheses for the position of the hexapods. These hypotheses were used to constrain the hexapods to different places in the Pancrustacea for the Bayes factor hypothesis tests.

Hypothesis name	Constraint
Null hypothesis	
H_0	Hexapods + Branchiopods
Alternative hypotheses	
H_1	Hexapods + Malacostracans
H_2	Hexapods + Cirripedes
H_3	Hexapods + Copepods
H_4	Hexapods + "Ostracods"
H_5	Hexapods + Malacostracans + Cirripedes
H_6	Hexapods + Malacostracans + Cirripedes + Copepods
H_7	Branchiopods + Malacostracans + Cirripedes + Copepods
H_8	Branchiopods + Malacostracans + Cirripedes + Copepods + "Ostracods"

constraints were run which separated the hexapods from the branchiopods by excluding them from various crustacean groupings (hypotheses H_7 and H_8).

The two taxa *Hutchinsoniella* and *Speleonectes* were removed from these analyses, as the position of *Hutchinsoniella* was very unstable between the previous analyses, whilst the strongly supported position of *Speleonectes* (as the sister-group to the Diplura) is likely to be artefactual (see section 3.5.1 for a detailed discussion). Otherwise, any constraints excluding these taxa from the hexapods could artificially reduce the likelihood of the analysis. To confirm that the removal of these taxa would not affect the recovered topology, the analysis was rerun with the two taxa removed using the R_2 model of the reduced dataset (as the analyses of this dataset appeared to converge better than those of the complete dataset) and modelling nucleotide substitutions with a GTR+G (as this model best fit the data in the previous analyses). This did not have any significant effect on the recovered topology (see figure 3.7) and the topology was reproduced by the alternative run: the two runs sampled similar log likelihood values (figure 3.8) and the split frequencies were low (table 3.11).

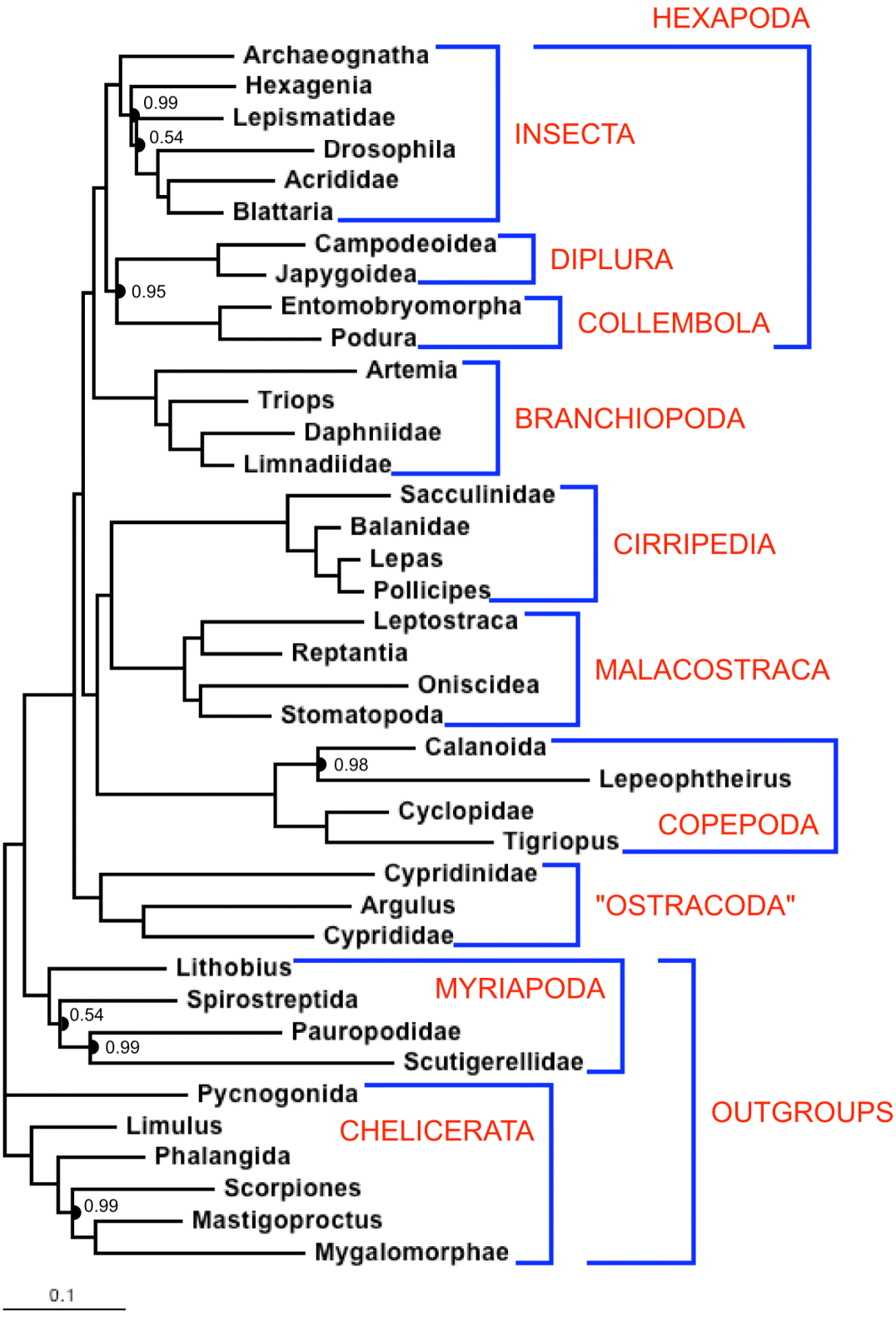


Figure 3.7. Consensus tree showing pancrustacean phylogeny with *Speleonectes* and *Hutchinsoniella* removed. The analysis used the reduced dataset (third codon position removed) and was analysed under the R_2 partitioning strategy and a GTR+G model of nucleotide substitution. The major pancrustacean groups are marked as are the outgroups. All nodes receive Bayesian posterior probability support values of 1.00, apart from the nodes marked with black half circles where the support value is indicated.

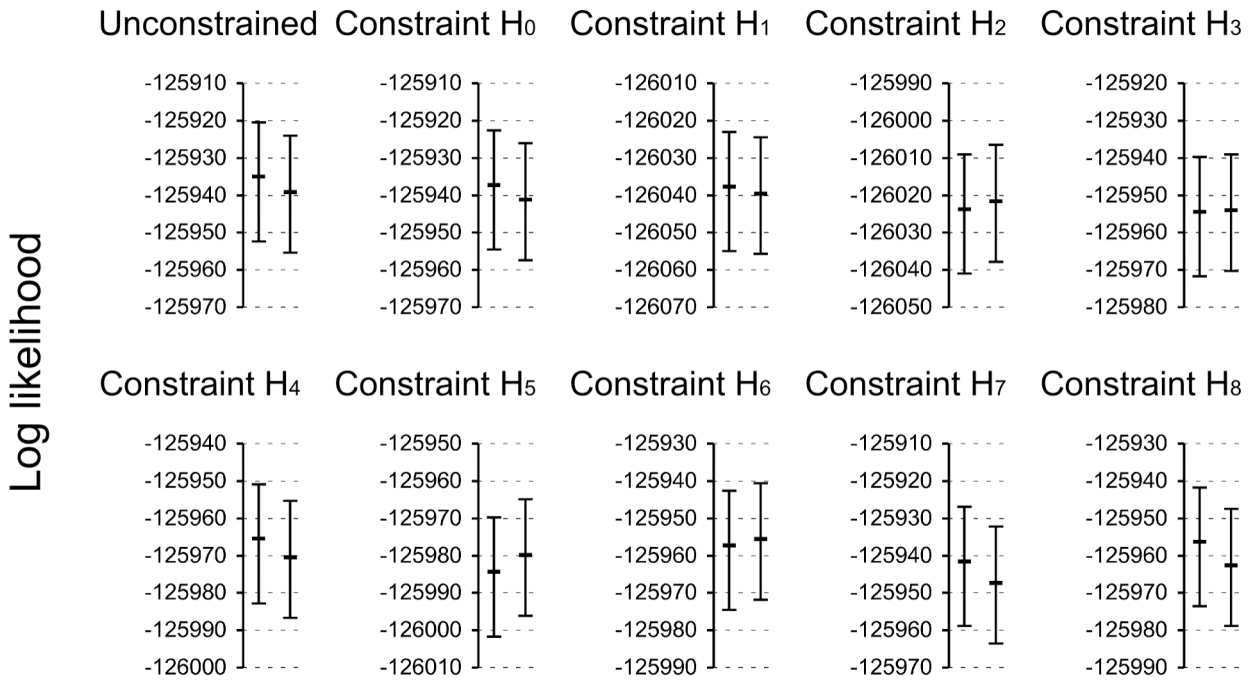


Figure 3.8. Comparisons of the distribution of log likelihoods for the two runs under the different topological constraints. For each topological constraint the arithmetic mean of the post burnin log likelihood values is plotted for the two runs. Bars show the range of the middle 95% of the log likelihood values. The different topological constraints are summarised in table 3.9. The distribution of log likelihoods for the two runs of the unconstrained analysis is also shown. The extent of the overlap of the two distributions gives an indication of whether they have sampled the same distribution of log likelihood values. For all constraints the two runs appear to have sampled similar distributions of log likelihoods.

Table 3.11. Split frequencies for the unconstrained and constrained runs without *Hutchinsoniella* and *Speleonectes*. All values are less than 0.05 indicating that the two runs sampled similar topologies.

Constraint	Split frequency
Unconstrained	0.007
H_0	0.018
H_1	0.023
H_2	0.032
H_3	0.014
H_4	0.015
H_5	0.006
H_6	0.041
H_7	0.010
H_8	0.015

Bayes factors

As in section 3.4.3 Bayes factor analyses were carried out using the preferred run for each analysis and the harmonic means were used as estimates for the marginal likelihoods. In addition, Bayes factors were also calculated using smoothed estimates of marginal likelihoods (Suchard, *et al*., 2005). For Bayes factor hypothesis tests to be reliable, the chain must have converged onto the posterior probability distribution. For all the different hypotheses, the two runs appeared to be sampling similar log likelihood values, and the split frequencies were generally low (in all cases below 0.05) (see figure 3.8 and table 3.11). If the runs had converged, pooling the data from the two runs for each constraint should not affect the Bayes factor values. Under this assumption of convergence, Bayes factors were also calculated after pooling the posterior distributions for the two runs of each hypothesis.

Preferred runs

Using the harmonic mean as an estimate of the marginal likelihood the Bayes factor analyses (summarised in table 3.12) found very strong support for the hexapod-branchiopod grouping (hypothesis H_0) over all of the other placements of the hexapods (Bayes factors >10). The only exception was the grouping of the branchiopods with the malacostracans, cirripedes and copepods to the exclusion of the hexapods (hypothesis H_7), where there was only strong support favouring the hexapod-branchiopod grouping with the Bayes factor falling below the cutoff of 10 (Bayes factor = 7.41).

Table 3.12. Bayes factors support for the hexapod-branchiopod grouping over other placements of the hexapods. The 2ln(BF) statistic gives the support for hypothesis H_0 over the specified alternative hypothesis. Different values of this statistic are given based on two different estimated for the marginal likelihood (harmonic mean and smoothed estimates) and using the posterior distribution of the preferred run alone and the two runs pooled together. Bayes factors giving ambiguous support (see section 2.2.8) are in bold.

Alternative hypothesis	Preferred run		Pooled runs	
	Harmonic mean	Smoothed marginal likelihood	Harmonic mean	Smoothed marginal likelihood
H_1	195.39	195.08	198.60	198.69
H_2	157.07	164.69	150.63	166.00
H_3	33.14	28.42	30.40	30.04
H_4	53.20	51.46	39.55	55.45
H_5	85.76	83.09	80.71	86.60
H_6	25.90	32.12	33.49	37.83
H_7	**7.41**	**5.73**	**-5.50**	11.30
H_8	23.19	33.23	32.33	41.26

Plotting 95% confidence intervals for these marginal likelihood estimates shows that some estimates have quite high variances (figure 3.9). In particular, there was a large amount of overlap in the distributions for hypotheses H_0 and H_7, suggesting that the Bayes factor value for the hexapod-branchiopod hypothesis could be an over- or underestimate.

Using smoothed estimates of marginal likelihoods Bayes factors the hypothesis H_0 was again favoured very strongly (Bayes factor >10) over all other hypotheses. The only exception was again hypothesis H_7 where the hexapod-branchiopod group received smaller positive support (Bayes factor = 5.73). Inspection of 95% confidence intervals shows that the smoothing gives tighter estimates of marginal likelihoods, suggesting that the Bayes factor values are more reliable (figure 3.9).

Pooled runs

Bayes factor estimates using the harmonic mean for pooled data give the same overall results as for the tests using the preferred runs: the hypothesis H_0 is favoured very strongly over all other hypotheses, with the exception of hypothesis H_7 (table 3.12). Here, in fact, there is positive support for the alternative hypothesis (Bayes factor = -5.50), although 95% confidence intervals again show that there is variability in the

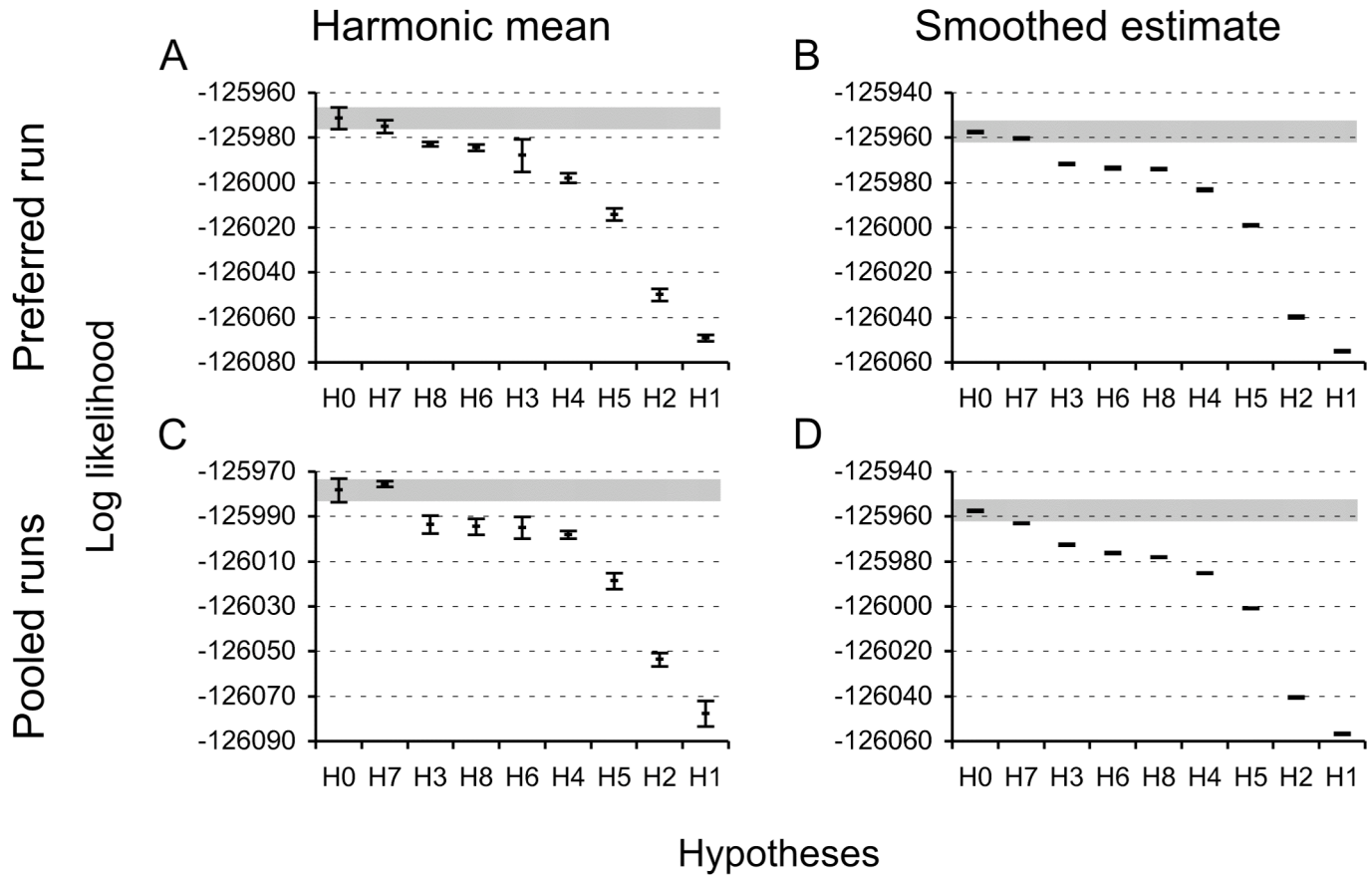


Figure 3.9. 95% confidence intervals for the estimates of the marginal likelihoods of the analyses run under the different topological constraints. Estimates of the marginal likelihood were calculated using the harmonic mean (A, B) or using smoothed estimates (C, D) and using only the preferred run (A, C) or the two runs for each constraint pooled (B, D). The alternative hypotheses are given in order of decreasing size of the marginal likelihood estimate. The grey bar represents the range of marginal likelihood estimates for which an alternative hypothesis would give an ambiguous result in the Bayes factors hypothesis test (±5 around the marginal likelihood estimate for hypothesis H_0). For all different estimates of marginal likelihood, only hypothesis H_7 fall within this ambiguous range, even when the variability in marginal likelihood estimates is accounted for. It is noteworthy that the smoothed estimates of marginal likelihoods (B, D) show less variability than the estimates using the harmonic mean (A, C).

estimates (figure 3.9). Calculating Bayes factors with smoothing again gives the same results (table 3.12), except that the hypothesis H_0 now also receives very strong support over the hypothesis H_7 (Bayes factor = 11.30). However, 95% confidence intervals show that there is a slight degree of variability in the estimates of the marginal likelihoods, and given that the Bayes factor is so close to the cutoff value (figure 3.9), it is difficult to argue confidently that the Bayes factor is greater than this value.

In summary, Bayes factor hypothesis tests give strong support to the hexapod-branchiopod grouping over all other hypotheses with the exception of excluding the hexapods from a grouping of the branchiopods with the malacostracans, cirripedes and copepods. Although there does appear to be some support for the hexapod-branchiopod grouping, the various different forms of the hypothesis test were ambiguous as to whether there was significant support over the alternative hypothesis.

3.5 Discussion

I have presented an analysis of pancrustacean phylogeny, specifically investigating the position of the insects. I have used a multigene dataset; the largest dataset yet used for a Bayesian analysis of pancrustacean phylogeny. I will now discuss the major phylogenetic results.

3.5.1 Pancrustacean phylogeny and the position of the insects

General comments

Due to various uncertainties in how best to model the data, I ran a number of analyses using different models of evolution. Whilst some models appeared to fit the data substantially better than others, all analyses recovered all the morphologically well-

supported taxa (Insecta, Diplura, Collembola, Branchiopoda, Malacostraca, Cirripedia, Copepoda, Myriapoda and Chelicerata) with strong posterior probability support. Therefore, even though there were many gaps in the dataset, there were no obvious artefactual placements of any taxa belonging to any of the well-established groups.

The position of Speleonectes *and* Hutchinsoniella

The only unexpected results related to the positions of the remipede *Speleonectes* and the cephalocarid *Hutchinsoniella*. All my analyses of nucleotide sequence found *Speleonectes* within the hexapods as the sister-group of the Diplura. Also, several analyses found strong support for *Hutchinsoniella* as the sister-group to this *Speleonectes*-Diplura group. This position, *within* the hexapods has not previously been suggested, and is unexpected. Both the remipedes and the cephalocarids are aquatic crustaceans with very distinctive bodyplans, so their placement within a highly tagmatised hexapodous terrestrial group is difficult to understand.

For *Hutchinsoniella* it is noteworthy that across all the different modelling strategies the position of this taxon was very variable, and its position varied between the preferred and alternative runs for several models. This is true to a lesser extent with *Speleonectes* where the position varied between the preferred and alternative runs for the amino acid coded analysis and the analysis using the R_2-*GTR* model. This instability between and within analyses gives reason to suspect that the placements of these two taxa within the hexapods may be artefactual. Interestingly, it has been shown that unstable taxa can artificially reduce the posterior probability support for a stable group by moving in and out of the group during one run of an MCMC (Dunn, *et al*., 2008). Importantly, this could potentially be the reason for the reduced support for the Hexapoda and Hexapoda + Branchiopoda as well as the Malacostraca + Cirripedia + Copepoda in the amino acid coded analysis and the analysis under the C_2-*GTR+G* model, as *Hutchinsoniella* and *Speleonectes* differ in their placements relative to these groups in the two runs of each analysis.

Based on analyses of nuclear protein coding genes, Regier *et al*. (2005) found the remipede and cepahlocarid taxa grouping with the hexapods and branchiopods,

although they found little support for a more resolved position within this group. They suggested that this hexapod-branchiopod-remipede-cephalocarid clade may be a "near-shore or marginal marine" group. Due to the difficulties I have shown for placing these taxa and the anomalous nature of their favoured positions, at the moment it is difficult to argue confidently for a close association with the hexapods. Investigations into potential biases in the signal are needed to see if any artefact in the data can be identified, which could be attracting these two taxa to a position with the diplurans.

The position of the insects

There was strong support for grouping the insects with the entognathous hexapod taxa: the diplurans and the collembolans. *Speleonectes* and *Hutchinsoniella* were also recovered within this grouping, but as these positions appear to be artefactual (as discussed above) there is support for a monophyletic Hexapoda. The recovery of the Hexapoda is in keeping with traditional views of arthropod phylogeny, and the position supported by the analyses based on nuclear datasets (Mallatt and Giribet, 2006, Mallatt, *et al*., 2004, Regier, *et al*., 2005), in contrast to the dual origins of the hexapods supported by the various mitochondrial studies (Cook, *et al*., 2005, Hassanin, 2006, Hassanin, *et al*., 2005, Lavrov, *et al*., 2004).

Across the different analyses there was also consistently strong support for a sister-group relationship between the hexapods and the branchiopods, a result that had previously been supported by the analyses of the nuclear protein coding genes and rRNAs. It is notable, however, that in the Bayes factor hypotheses tests, the support for this hexapod-branchiopod group over a topology where the branchiopods were constrained to group with the malacostracans, cirripedes and copepods fell below the critical cutoff value for Bayes factors of 10. This was the only alternative hypothesis over which the hexapod-branchiopod grouping did not receive very strong support.

Therefore, whilst the posterior probability support values for the various taxon bipartitions of the most suitable modelling strategies provide overwhelming evidence for a hexapod-branchiopod grouping, Bayes factor hypothesis tests give weaker support, providing some support to an alternative position of the branchiopods. Interestingly, the

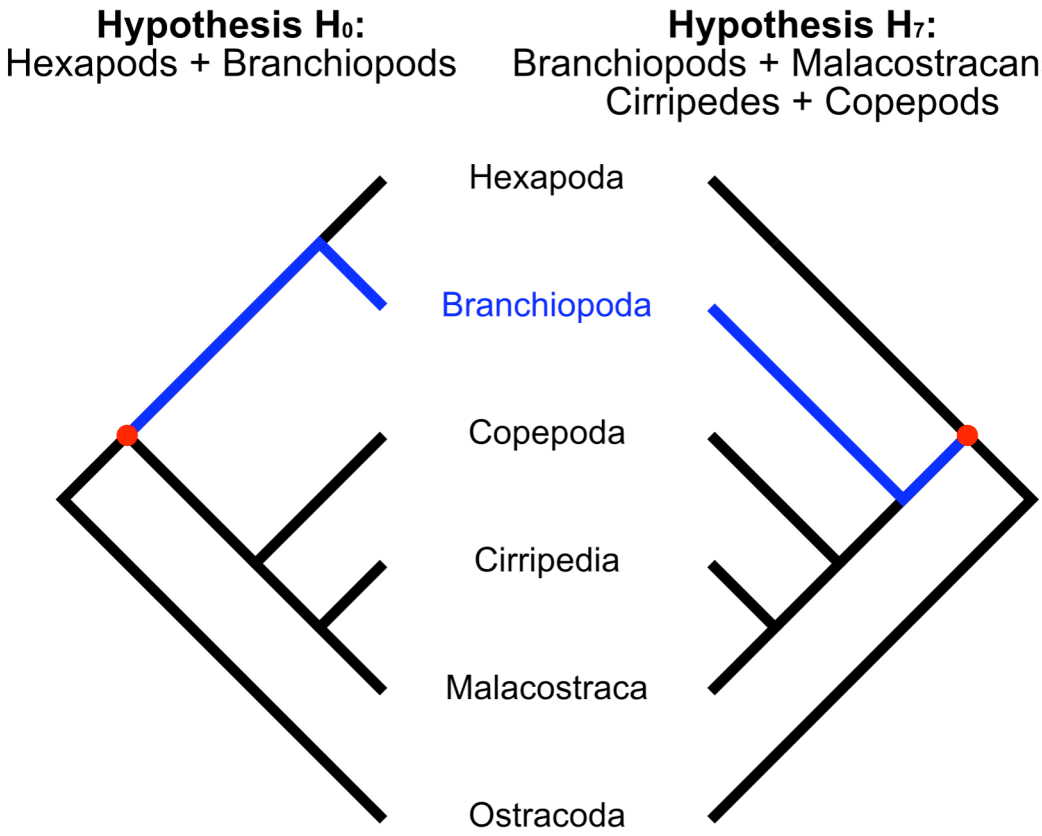


Figure 3.10. Alternative positions for the branchiopods under the hexapods + branchiopods constraint (hypothesis H_0) and the branchiopods + malacostracans + cirripedes + copepods constraint (hypothesis H_7). Schematics of the consensus trees for the analyses run under the two constraints show that the alternative positions for the branchiopods only require the movement of the branchiopod branch (blue) across one node (red circle) in an otherwise stable topology.

consensus trees for the two analyses show that the only difference between the two hypotheses is in the movement of the branchiopods across one node in an otherwise stable tree (figure 3.10). This resembles a soft polytomy, suggesting that there may be a weak signal at this node rather than the branchiopods being attracted to very different places in the tree.

A hexapod-malacostracan relationship is strongly rejected

The different analyses found strong support for a number of other features of pancrustacean phylogeny. Of particular interest was the position of the malacostracans. This group of crustaceans had previously been implicated in the origins of the hexapods, most recently on the basis of shared features of the brain (for more discussion of this see section 3.5.2). One of the strongest results of all my analyses, however, was

the grouping of the malacostracans with the cirripedes. This was recovered with high posterior probability support across all different modelling strategies. Additionally, the different analyses also found strong support for grouping this malacostracan-cirripede clade with the copepods. Although this grouping was not as strongly supported as the malacostracan-cirripede group, it does place further taxonomic distance between the hexapods and malacostracans. Also, the in the Bayes factor hytpothesis tests, the hypothesis constraining the hexapods with malacostracans was one of the least favoured hypotheses. These results strongly argue against any close relationship between the hexapods and malacostracans.

Support values

It is notable that several of these groups identified within the Pancrustacea are supported by high posterior probabilities. However, it has often been argued that posterior probabilities give over-confidence in results (for example Huelsenbeck, *et al*., 2002). It is possible that the high level of support is an artefact. An alternative method of assessing support is bootstrapping, which tends to give more conservative estimates. However, programmes for bootstrapping do not implement the doublet model used in these analyses, and so for now it is not possible to provide this alternative measure of support. Therefore, it is sensible to view the high support values with caution. The ambiguity in the support for the hexapod-branchiopod group identified in the hypothesis tests may in fact be an accurate reflection of an actual lower level of support for the group. In the absence of bootstrapping, such hypothesis tests may be one way of testing the support for important nodes in the tree.

3.5.2 Comparison to previous analyses

One of the motivating factors behind this study was to address the uncertainties in pancrustacean phylogeny between the previous analyses of the smaller datasets, by using a combined analysis.

Analyses of nuclear datasets

As has already been described, the combined analysis largely resembles the smaller analyses based on the two nuclear datasets: the protein coding genes EF-1α, Pol II and EF-2 by Regier *et al*. (2005) and the 18S and 28S rRNAs by Mallatt and Giribet (2006). One of the important differences between these two studies was the position of the copepods, with the protein coding genes supporting a position with the malacostracans and cirripedes, whilst the rRNAs gave some support for a position as sister-group to the hexapods. My analysis gives strong support to the former of these two hypotheses, with the copepods repeatedly grouping with the malacostracans and cirripedes with posterior probabilities close to 1.00 in the best-supported analyses. Furthermore, the Bayes factor hypothesis tests also found very strong support for the hexapod-branchiopod sister-grouping over a hexapod-copepod sister-grouping. Therefore, this combined dataset finds no support for grouping the hexapods with the copepods. It seems that the hexapod-copepod sister grouping supported by the rRNAs is, as Mallatt and Giribet (2006) suggest, an analytical artefact perhaps relating to the divergent sequence of the single copepod represented (Mallatt*, et al*., 2004).

Analyses of mitochondrial genomes

The analyses presented here recover a different phylogeny for the Pancrustacea to that favoured by the several analyses based on mitochondrial genomes (Carapelli*, et al*., 2007, Cook*, et al*., 2005, Hassanin, 2006, Hassanin*, et al*., 2005, Lavrov*, et al*., 2004, Nardi*, et al*., 2003). These mitochondrial analyses recovered paraphyletic hexapods, with the insects grouping with the malacostracans and branchiopods to the exclusion of the collembolans and the maxillopod crustaceans. However, as discussed above, it bears strong resemblance to the topologies based on the individual nuclear datasets.

It is possible that this is a genuine signal from the combined nuclear and mitochondrial datasets. The mitochondrial genes are fast evolving and so there may be less signal in the data to conflict with the signal of the nuclear genes; the low proportion of invariant sites compared to the nuclear genes (see table 3.1) are suggestive that this may be the case. When analysed alone this lack of signal may lead to an artefactual topology. It is

noteworthy that the mitochondrial analyses did not receive strong support from non-parametric bootstrapping and alternative hypotheses could not be rejected (Carapelli, *et al.*, 2007, Cook, *et al.*, 2005, Hassanin, 2006, Hassanin, *et al.*, 2005, Lavrov, *et al.*, 2004, Nardi, *et al.*, 2003). In contrast the topologies recovered by the analyses of the nuclear datasets were generally strongly supported by non-parametric bootstrapping (Mallatt and Giribet, 2006, Regier, *et al.*, 2005). A gene by gene investigations of the phylogenetic content of each nuclear and mitochondrial gene (for example through likelihood mapping or saturation plots) could give an indication of the signal in the two datasets. If there is a weaker signal in the mitochondrial datasets than the nuclear datasets, then the resemblance of my combined analysis to the nuclear analyses is not unexpected.

However, the resemblance of my multigene analyses to the nuclear gene analyses could be an artefact resulting from the composition of the dataset. Out of the 16370 nucleotide sites (including third positions), 9464 are from nuclear genes (18S, 28S, EF-1a, EF-2, PolII and H3) whilst 6906 are from mitochondrial genes. Furthermore, the mitochondrial sites in the dataset have more incomplete taxa, being only 62.5% complete compared to 81.9% for nuclear genes. Therefore, the signal of the nuclear genes may have swamped the signal from the mitochondrial genes. It is difficult to see how this could be tested without sequencing more mitochondrial genomes, although perhaps analysing a reduced dataset containing fewer nuclear sites could give an indication as to whether the nuclear signal had obscured the mitochondrial signal.

Implications for arthropod neurobiology

Apart from the molecular phylogenetic analyses, the most notable other source of data that has been put forward to support the Pancrustacea comes from neurobiology. Of particular interest are analyses based on brain morphology. Based on proposed shared derived features of the optic lobes it was suggested that the insects and malacostracans form a clade (Harzsch, 2002, Sinakevitch, *et al.*, 2003). Specifically, pterygote insects and decapods have three optic neuropils connected by chiasmatising fibres, whilst "entomostracan" crustaceans (represented by the branchiopods) have only two neuropils connected by parallel fibres.

Our results strongly reject this hypothesis. In fact, constraining the hexapods to group with the malacostracans was one of the most strongly rejected hypotheses. Our results would mean that these similarities in brain structure evolved convergently. There is some evidence that there has been at least some convergence in the optic lobes of pterygotes and decapods. In the basal members of the insects and malacostracans, namely the Archaeognatha and the Phyllocarida respectively, the optic lobes only consist of two neuropils (although these are connected by chiasmatising fibres) (Sinakevitch, *et al*., 2003). Therefore, there is at least some level of variability in brain structure within the insects and malacostracans, and so it is not unreasonable to suggest that their shared brain structure could have evolved convergently.

3.5.3 Methodological considerations

The primary goal of the analyses described in this chapter was to resolve pancrustacean phylogeny and the position of the insects. However, in running the analyses, a number of different phylogenetic methods were used. Various considerations relating to the use of these methods are worthy of a brief discussion.

Modelling strategies and convergence

A range of different treatments of codon position and substitution model were run, as the most appropriate model to analyse the data was not obvious on the basis of any *a priori* evidence. Whilst there was a large amount of topological agreement between the different analyses, it was clear the different models behaved differently. This was most obviously seen in topological differences between the consensus trees or in differences in the posterior probabilities for various groupings. There also seemed to be differences in how well the MCMCs converged on the posterior distribution.

There is a suggestion that the MCMC had more trouble converging on the posterior distribution when there was a poorer fit of the model to the data. This was seen most clearly when the effect of different models of nucleotide substitution was investigated (section 3.4.4). As the fit of the model to the data was worse when the GTR or HKY

substitution models were used without gamma distributions – as shown by Bayes factors and the estimates of the AIC and BIC – the topological convergence between the runs fell and the runs did not plateau with the same distribution of log likelihoods.

Additionally, in the investigations of different treatments of the codon positions (section 3.4.3) there appeared to be more problems with convergence when the third codon position was included. In these modelling strategies (C_1, C_2 and C_3) the two runs sampled different log likelihood distributions and for two of the models (C_2 and C_3) the topological convergence between the runs was poorer. These problems were not apparent for the two runs without the third codon positions. The investigations of nucleotide composition (section 3.4.2) suggested that there was a large degree of heterogeneity at the third position for both nuclear and mitochondrial genes. Perhaps there were difficulties in modelling this heterogeneity and these difficulties led to problems with convergence.

These inferences are all based on how well two runs converged for each model. To make any strong statements on how the fit of a model affects convergence on the posterior distribution more runs would be needed. However, the results presented here give some potentially interesting insights.

Bayes factors: Favoured models

Bayes factors were used extensively as a means of choosing between different models and different phylogenetic hypotheses. This was partly due to their ease of use, as at the simplest level all that is needed is the harmonic mean of the sampled log likelihood values, and also because the use of likelihood based methods was not possible as the preferred models could not be implemented in likelihood packages. However, a number of features relating to the use of Bayes factors became apparent, which warrant discussion.

There have been questions about how Bayes factors respond to the addition of parameters, with several studies suggesting that Bayes factors tend to support parameter rich models. It is therefore important to note that in my analyses increasing the number

of parameters did not necessarily improve the fit of the model to the data. In my investigation of different nucleotide substitution models whilst the best modelling strategy in terms of models was the most parameter rich (GTR+G), the HKY+G fit the data better than the more parameter rich GTR. The heterogeneity between sites provided by the gamma distribution appears to be more important than the heterogeneity in the substitution process provided by the GTR model.

Bayes factors: Potential problems

My use of Bayes factors also highlighted various important features that need to be considered when using this method. Bayes factors require an estimate of the marginal likelihood: most commonly the harmonic mean is used. However, as was shown for the hypothesis tests (section 3.4.6), constructing 95% confidence intervals shows that these estimates can have a high variance. Whilst this may not be a problem if the Bayes factor estimates are orders of magnitude greater than the cutoff of 10 (as in sections 3.4.3 and 3.4.4) it could be a potential problem when Bayes factors are smaller (as in section 3.4.6). It is, therefore, important to consider this variability and use a possibly less variable estimate of the marginal likelihood, such as using smoothing to be confident in the results of the tests, or at least calculate the variance.

Perhaps more importantly, for a Bayes factor to be reliable, the run must have converged on the posterior distribution. Using split frequencies and examining the distributions of log likelihoods I judged that my different runs had converged on the posterior distribution. Also, pooling the two runs did not affect the overall result: the hexapod-branchiopod grouping was favoured over all other hypotheses other than the branchiopod-malacotsracan-cirripede-copepod grouping. Despite this, some hypotheses did change their Bayes factor values by around 10 for example the hypothesis grouping the hexapods with the ostracods (hypothesis H_4). Whilst in this situation the overall result was not affected, as the Bayes factor was considerably greater than 10, in different circumstances such a change could have been significant. Therefore, it is important to consider potential problems with convergence in order to have confidence in Bayes factors.

3.6 Conclusions

I set out to analyse pancrustacean phylogeny and in particular the position of the insects, assembling the largest yet multigene dataset with a broad representation of hexapod and crustacean taxa. Importantly, my analyses have provided strong support for a phylogeny of the Pancrustacea with a monophyletic Hexapoda and a sister-group relationship between these hexapods and the branchiopod crustaceans. Using Bayes factor hypothesis tests I have been able to reject a number of alternative hypotheses for sister-groups to the hexapods that had been proposed in the literature, such as the hexapod-copepod sister-group relationship supported by the analyses of 18S and 28S rRNAs and the hexapod-malacostracan sister-group supported by brain morphology. This emerging picture of pancrustacean phylogeny will provide a framework in which to ask questions about insect bodyplan evolution and to infer developmental changes underlying the morphological transitions. In the following chapters I will now address one such transition, namely the evolution of the intercalary segment of the insect head, in particular investigating how the segment develops in the insects.

Chapter 4:

The *Drosophila* intercalary segment and the affinity of the hypopharyngeal lobes

The results described in this chapter are currently in press: Economou, A. D. and Telford, M. J. Comparative gene expression in the heads of *Drosophila melanogaster* and *Tribolium castaneum* and the segmental affinity of the *Drosophila* hypopharyngeal lobes. Evol. Dev. In Press.

4.1 Summary

In this chapter I address the issue of what constitutes the intercalary segment in the model organism *Drosophila melanogaster*. The *Drosophila* embryonic head has a pair of structures lying behind the stomodeum known as the hypopharyngeal lobes. Traditionally they have been seen as part of the intercalary segment. More recent work looking at the position of the lobes relative to various marker genes has been somewhat equivocal: segment polarity gene expression has been used to argue for a mandibular affinity of these lobes, whilst the expression of the anterior-most hox gene *labial* (*lab*) has supported an intercalary affinity. I have addressed the question of the segmental affinity of the hypopharyngeal lobes by conducting a detailed comparison of gene expression patterns between *Drosophila* and the red flour beetle *Tribolium castaneum*, in which the intercalary segment is unambiguously marked out by the expression of *lab*. I demonstrate that there is a large degree of conservation in gene expression patterns between *Drosophila* and *Tribolium*, and this argues against an intercalary segment affinity for the hypopharyngeal lobes. The lobes appear to be largely mandibular in

origin, although some gene expression attributed to them appears to be associated with the stomodeum. I propose that the difficulties in interpreting the *Drosophila* head result from a topological shift in the *Drosophila* embryonic head, associated with the derived process of head involution.

4.2 Introduction

Having addressed the issue of pancrustacean phylogeny and the position of the insects, I now concentrate on the development of the intercalary segment. As was illustrated in chapter 1, only a limited amount is known about intercalary segment development and the majority of what is known comes from the vast literature from the model organism *Drosophila melanogaster*. However, there are difficulties in extrapolating from what is known in the fly to other insects, largely because there is a lack of consensus as to what constitutes the *Drosophila* intercalary segment. In this chapter I address this issue.

The difficulty in interpreting the *Drosophila* intercalary segment stems from the highly derived mode of head embryogenesis seen in the fly. As with other cyclorrhaphan flies, *Drosophila* head embryogenesis is notable for the process of head involution. The head segments pass through the stomodeum (for a detailed description see Turner and Mahowald, 1979) giving rise to the acephalic maggot larva. This larva possesses an atypical set of head structures, with the cells giving rise to the typical insect head of the adult being set aside as imaginal discs (Younossi-Hartenstein, *et al*., 1993). The structures of the larval head have proved very difficult to homologise with the components of the canonical insect head (Jürgens, *et al*., 1986).

Prior to involution the fly embryo does not bud out head appendages as insects typically do; rather the embryonic head has the appearance of a series of lobes (figure 4.1 A). During germband retraction a further set of lobes form immediately posterior to the stomodeum: the hypopharyngeal lobes (Turner and Mahowald, 1979). These have

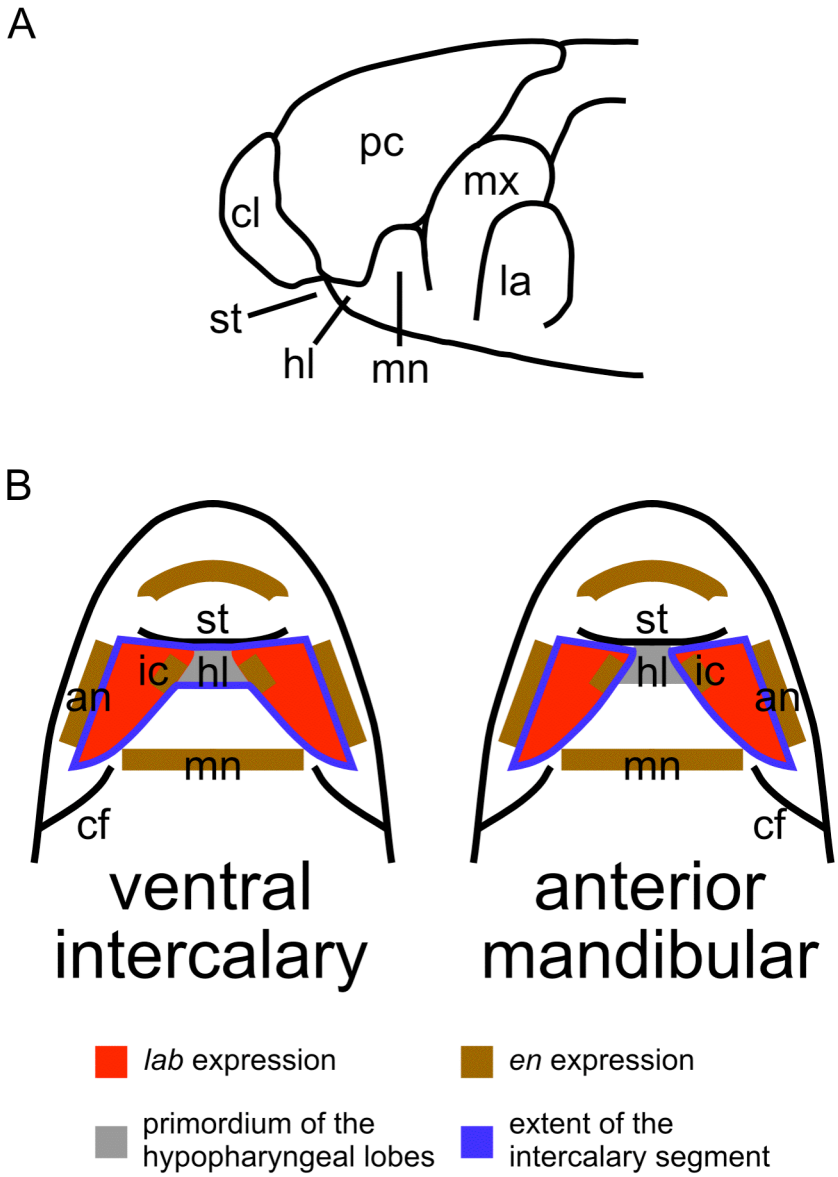


Figure 4.1. The *Drosophila* head and the hypopharyngeal lobes. (A) Schematic of the head of a *Drosophila* stage 11 embryo showing the series of lobes that make up the embryonic head, and the primordium of the hypopharyngeal lobes posterior to the stomodeum. (B) Different interpretations of the hypopharyngeal primordium as the ventral intercalary segment or the anterior mandibular segment. The hypopharyngeal primordium (grey) has been argued to be ventral intercalary as it lies medial to the *lab* domains (red) (Rogers and Kaufman, 1997). It has also been interpreted as mandibular as it lies posterior to the segment polarity gene stripes of the intercalary segment (*en* expression is shown in brown) (Diederich, *et al*., 1989, Diederich, *et al*., 1991, Mohler, *et al*., 1995, Seecoomar, *et al*., 2000). The extent of the intercalary segment in the two interpretations in marked out in blue. The stomodeum and cephalic furrow are also marked. The expression of *lab* and *en* is based on Mahaffey *et al*. (1989). an, antennal; cf, cephalic furrow; cl, clypeolabral; hl, primordium of the hypopharyngeal lobes; ic, intercalary; la, labial; mx, maxillary; mn, mandibular; pc, procephalic, st, stomodeum.

traditionally been interpreted as part of the intercalary segment (Rogers and Kaufman, 1997), largely due to their position posterior to the stomodeum where the intercalary segment is found in other insects, but also by comparison to the paired lobes, often called hypopharyngeal lobes (or *hypopharynxhöcker*) known to arise from the intercalary segment in numerous other insect groups (Roonwal, 1937, Wolff and Scholtz, 2006).

Based on the interpretation of the *Drosophila* hypopharyngeal lobes as intercalary derivatives, three genes have been implicated in patterning the segment: *cap'n'collar* (*cnc*), *knot* (*kn*) (synonymous with *collier* (*col*)) and *crocodile* (*croc*). *cnc* (a leucine zipper transcription factor) is expressed posterior to the stomodeum in the developing hypopharyngeal lobes (Mohler, *et al.*, 1991), and is required for the differentiation of the posterior pharyngeal wall (Mohler, *et al.*, 1995) – a structure mapped to the hypopharyngeal lobes by Jürgens *et al.* (1986). *kn* (a COE transcription factor) is expressed along the intercalary-mandibular boundary and appears to be required for the expression of *cnc* in the hypopharyngeal lobes, as well as for the expression of the intercalary segment polarity genes (Crozatier, *et al.*, 1999, Seecoomar, *et al.*, 2000). *croc* (a fork head transcription factor) is also expressed posterior to the stomodeum in this hypopharyngeal region and is required for the formation of the posterior pharyngeal wall (Häcker, *et al.*, 1995), although it is not clear from published literature how its expression and function fit in with that of *cnc* and *kn*.

However, this interpretation of the *Drosophila* head has been questioned. Diederich *et al.* (1989, 1991) argue that there is no association between the hypopharyngeal lobes and any *en* expression, proposing that they therefore belong to the anterior of the mandibular segment. Similarly, Mohler *et al.* (1995) and Seecoomar *et al.* (2000) argue that expression of two of the genes that are required to pattern the lobes (*cnc* and *kn* respectively) lies posterior to the intercalary *hedgehog* (*hh*) stripes and should therefore be considered part of the mandibular segment.

A problem with these arguments, as Rogers and Kaufman (1997) point out, is that there is a large gap separating the stripes of intercalary segment polarity gene expression and this is a derived feature of *Drosophila*. Other insects do not show such a large

separation of their *engrailed* (*en*) stripes as *Drosophila*, and Rogers and Kaufman (1997) argue that *en* is a poor marker for the *Drosophila* intercalary segment. Rather, they propose that *lab* is an appropriate marker for the intercalary segment as it is expressed throughout the segment in other insects. They argue that this is also the case in the early *Drosophila* embryo as is shown by Diederich *et al.* (1989); *lab* is expressed throughout the intercalary segment before fading from the ventral regions that give rise to the hypopharyngeal lobes. The two different interpretations of the intercalary segment are shown in figure 4.1 B.

In other insects such as the red flour beetle, *Tribolium castaneum*, where *lab* expression is seen in a continuous domain that unambiguously marks the intercalary segment (Nie, *et al.*, 2001) these difficulties in identifying the intercalary segment do not exist. There is no suggestion of fading from the ventral part of the segment as in *Drosophila*. If the genes expressed in the *Drosophila* hypopharyngeal lobes can be demonstrated to have conserved patterns of expression in *Tribolium*, then it should prove relatively easy to determine whether any given pattern of gene expression belongs to the intercalary segment or to the mandibular segment.

I have identified and cloned partial cDNAs of *Tribolium* orthologues of the three genes – *cnc*, *croc* and *kn* – that are involved in patterning the *Drosophila* hypopharyngeal lobes. I examined the expression patterns of these genes in *Tribolium* embryos and compared them to what is seen in *Drosophila*. Where necessary, I also re-examined the expression patterns in *Drosophila* to facilitate detailed comparisons between the two insects, through time. I used double *in situ* hybridisations in both *Tribolium* and *Drosophila* to compare the expression patterns of the genes of interest to *lab* and to each other. This approach allows me to rule out an intercalary segment affinity for the fly hypopharyngeal lobes and a role in intercalary segment development for the genes that pattern the lobes. I also propose an explanation for what underlies some of the peculiarities of gene expression in the *Drosophila* embryonic head.

4.3 Materials and Methods

Tribolium and *Drosophila* stocks were maintained as described in section 2.3.1, and embryos were collected and fixed as described in sections 2.3.2, 2.3.3 and 2.3.9. *Tribolium* orthologues of the *Drosophila* genes were identified in the *Tribolium* genome (see section 2.3.5; accession numbers for *Drosophila* query sequences are given in appendix 1 table A1.3) and partial cDNAs were amplified by PCR from *Tribolium* cDNA (primer sequences are given in appendix 2, table A2.2) and cloned (see sections 2.3.4 and 2.3.6). For *cnc*, the B isoform was used as the query sequence in the BLAST search as this is the isoform that has been implicated in head development (Veraksa, *et al*., 2000). *Tc-lab* was cloned using primers designed against the published sequence (Nie, *et al*., 2001) and a clone of *Tc-en* was kindly donated by Dr Andrew Peel. For *Drosophila* genes, complete cDNAs were ordered from the *Drosophila* Gene Collection as described in section 2.3.7 (clone names are given in appendix 3, table A3.1). *In situ* hybridisation was carried out as described in section 2.3.10 using DIG labelled probes (see section 2.3.8). Double *in situ* hybridisation was carried out as described in section 2.3.11 using DIG labelled probes and fluorescein labelled probes (see section 2.3.8). Embryos were prepared and imaged as described in section 2.3.13.

4.4 Results

Unambiguous orthologues of *cnc*, *croc* and *kn* were identified in the *Tribolium* genome. I first describe the expression of these three genes in *Tribolium*, comparing them to the *Drosophila* expression patterns. Where necessary, I present a re-examination of the *Drosophila* pattern.

4.4.1 Expression of cap'n'collar *orthologues in* Tribolium *and* Drosophila

Tc-cnc shows many similarities in its expression to its *Drosophila* ortholog. There is an anterior domain of expression, which resolves to the labrum, and a posterior domain, which resolves to the mandibular segment (figure 4.2 A, C, E). This resembles the "cap" and "collar" of expression seen in *Drosophila* (Mohler, *et al*., 1991) (although the exact segmental affinity of the "collar" in *Drosophila* is uncertain given the ambiguities over the hypopharyngeal lobes). In addition, during late germband extension, the anterior domain in *Tribolium* extends posteriorly to form a ring around the stomodeum (figure 4.2 C, E). My re-examination of *Drosophila* shows a very similar expression domain posterior to the stomodeum, which also appears during germband extension (figure 4.2 D). The major difference between *Tribolium* and *Drosophila* is that, whilst there is a gap between this stomodeal domain and the "collar" of expression in *Tribolium*, these two domains abut in *Drosophila* (compare figure 4.2 E with figure 4.2 D). In the germband extended *Drosophila* embryo, this results in a continuous domain of expression from the mandibular lobes extending throughout the hypopharyngeal primordium (figure 4.2 F) as described in Mohler *et al*. (1991).

There are also similarities in the early expression of *cnc*. As in *Drosophila*, there is no expression in the prospective mesoderm of *Tribolium* (figure 4.3 B) (compare to Mohler, *et al*., 1991). However, there are also some subtle differences in the early expression. In *Tribolium*, the anterior domain is initially seen as a pair of expression domains at the anterior of the embryo (figure 4.3 C). The single continuous anterior expression domain is only seen later (figure 4.3 D). In contrast, no initial pair of expression domains has been described for *Drosophila* (Mohler, *et al*., 1991). Also, whilst in *Drosophila* the "cap" of expression appears before the "collar" (Mohler, *et al*., 1991), in *Tribolium* it is the other way round (figure 4.3 A).

4.4.2 Expression of crocodile *orthologues in* Tribolium *and* Drosophila

Tc-croc has a dynamic expression pattern, which recapitulates many of the features described for its *Drosophila* orthologue (Häcker, *et al*., 1995). Expression is first seen

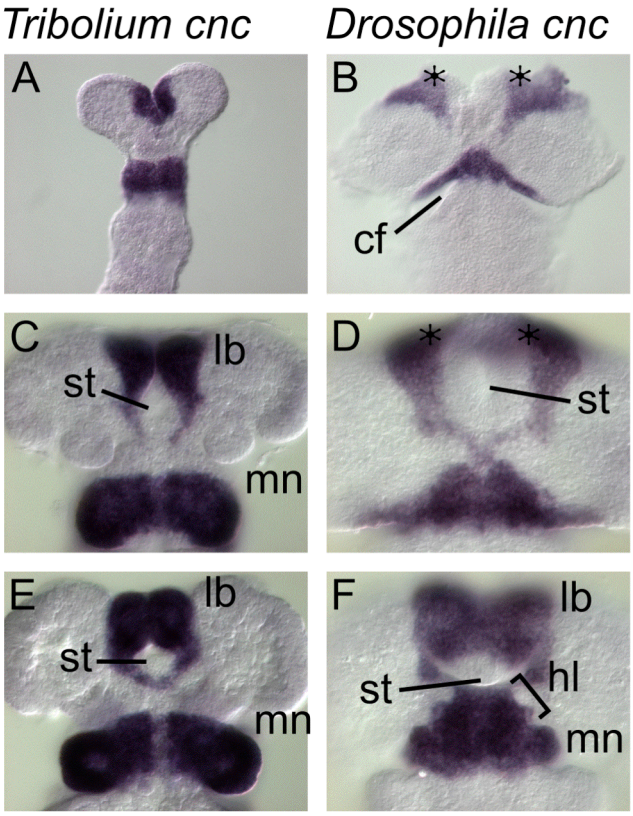


Figure 4.2. Expression of *cap'n'collar* orthologues in *Tribolium* and *Drosophila*. (A, C, E) *Tribolium*, (B, D, F) *Drosophila*. Ventral views, oriented with anterior up. Nomarski images. In both *Tribolium* and *Drosophila*, *cnc* orthologues are first seen in an anterior domain and a more posterior band; posterior to the head lobes in an early germband extending *Tribolium* embryo (A) and anterior to the cephalic furrow in an early germband extending (stage 7) *Drosophila* embryo (B). In *Tribolium*, the anterior domain resolves to the labrum, and the posterior domain to the mandibular segment, as seen in a late germband extending embryo (C). In addition, the anterior domain in *Tribolium* appears to extend posteriorly around the forming stomodeum (C), and by the end of germband extension, this domain forms a ring around the stomodeum (E). Similarly in *Drosophila*, by late germband extension (stage 9) (D), expression is seen extending from the posterior of the early "cap" of expression, to form a ring around the stomodeum. Unlike *Tribolium*, this domain abuts the "collar" of expression. By the end of germband extension (stage 11) (F), when the *Drosophila* head lobes have formed, expression has resolved to the labrum, the mandibular segment (as seen by expression throughout the mandibular lobes), and the primordium of the hypopharyngeal lobes (as seen by expression extending from the mandibular lobes to the posterior edge of the stomodeum). Note that the "cap" of expression appears broken into two (asterisk in B and D), as the procephalic lobe was split dorsally to allow the embryo to be flattened. cf, cephalic furrow; hl, primordium of the hypopharyngeal lobes; lb, labrum; mn, mandibular lobes; st, stomodeum.

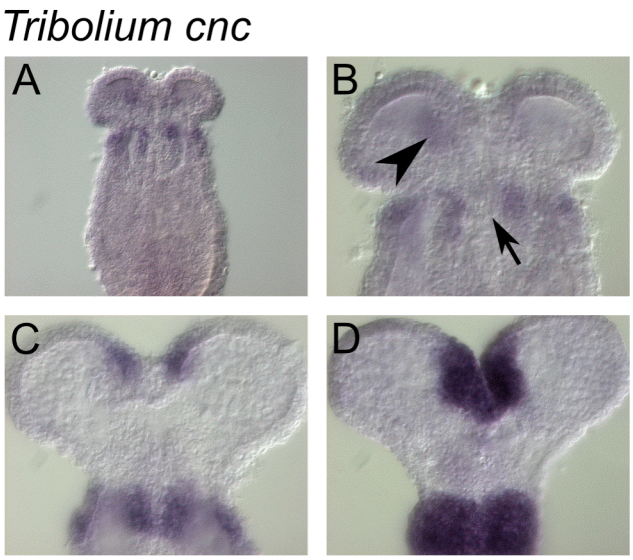


Figure 4.3. Early expression of *Tribolium cap'n'collar*. Ventral views, oriented with anterior up. Nomarski images. *Tc-cnc* is first seen as a distinct band of expression immediately posterior to the head lobes in the gastrulating embryo (A). A close up of the same embryo (B) shows that this band of expression does not extend into the forming mesoderm (arrow). There is also potential faint expression at the anterior of the embryo (arrowhead). By early germband extension a clear posterior "collar" of expression can be seen (C). The faint anterior expression can now be seen as a clear pair of domains. Later in germband extension (D) the "collar" is still visible and the anterior pair of domains have joined together to form a single "cap" of expression at the anterior of the embryo (D).

in an early anterior domain (figure 4.4 A) before retracting from the anterior-most region which appears to correspond to the prospective stomodeum (figure 4.4 C). Expression then further reduces, with transcripts not seen in the forming labrum when it is first clearly visible (figure 4.4 E). This leaves an expression domain immediately posterior and lateral to the stomodeum in the late germband extending embryo (figure 4.4 G). Detailed comparison of the *Tribolium* expression pattern (figure 4.4 A, C, E, G) with *Drosophila* (figure 4.4 B, D, F, H) show how consistent the similarities are. Given the proposed role of *croc* in patterning the *Drosophila* intercalary segment (Häcker*, et al.*, 1995), it is noteworthy that expression is not extensive in the *Drosophila* hypopharyngeal primordium (figure 4.4 H).

Although there are many striking similarities in the expression patterns of *Drosophila* and *Tribolium croc*, there are some subtle differences in the modulations. Whilst in both *Drosophila* and *Tribolium croc* orthologues are expressed in the mesoderm early in embryogenesis (figure 4.5 C and D), in *Drosophila* this early mesodermal expression of *croc* fades, and the expression in the ectoderm is unconnected ventrally (figure 4.5 F). The domain posterior to the stomodeum of older embryos is seen later (figure 4.5 H). This is not the case in *Tribolium*. After the posterior mesodermal expression of *Tc-croc* has faded (figure 4.5 E) there is no obvious gap in the expression of in the ventral ectoderm.

4.4.3 Expression of Tribolium knot

Early *Tc-knot* expression strongly resembles what is seen in its *Drosophila* orthologue as described by Crozatier *et al*. (1996, 1999). Transcripts first accumulate at the posterior of the procephalon behind the *Tribolium* head lobes (figure 4.6 A), which compares with expression immediately anterior to the *Drosophila* cephalic furrow (Crozatier*, et al.*, 1996). As with *Drosophila*, this early blastodermal expression is bounded posteriorly by the *Tc-hh* expressing cells of the mandibular parasegment boundary (parasegment 0; figure 4.6 B). Anteriorly, *Tc-kn* abuts the antennal domain of *Tc-hh* (figure 4.6 B). This also appears conserved with *Drosophila*, although Crozatier *et al*. (1999) describe this anterior *hh* domain as an asegmental cephalic domain.

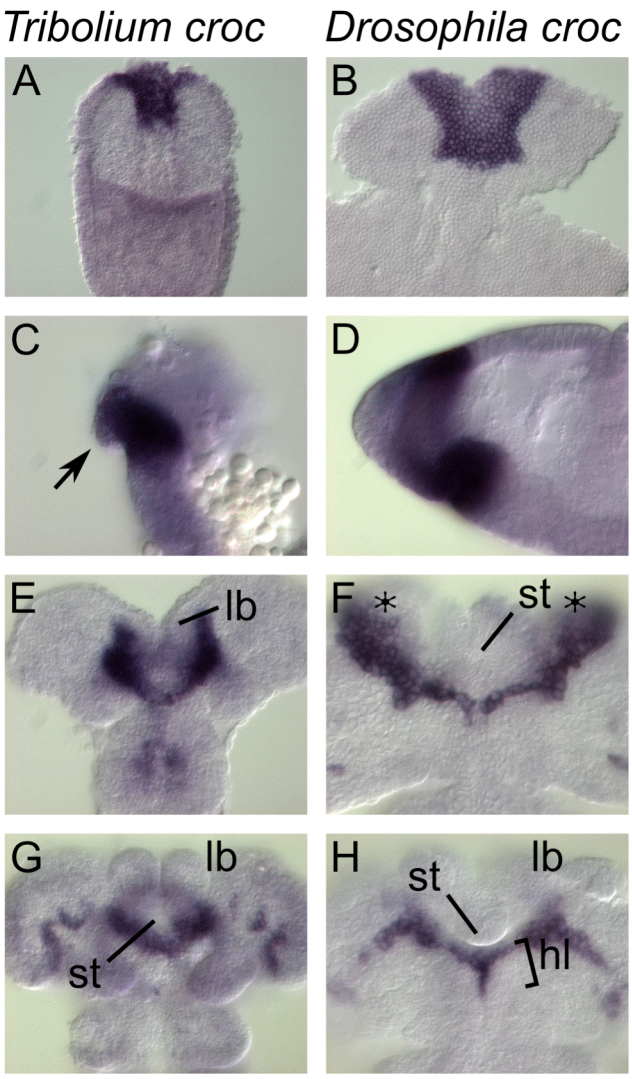


Figure 4.4. Similarities in the expression of *crocodile* orthologues in *Tribolium* and *Drosophila*. (A, C, E, G) *Tribolium*, (B, D, F, H) *Drosophila*. (A, B, E-H) Ventral view of embryos, oriented with anterior up. (C, D) Lateral view of embryo, oriented with anterior left. Nomarski images. In both *Tribolium* and *Drosophila*, expression is first detected at the anterior of the early embryo, as can be seen in the gastrulating *Tribolium* embryo (A) and the cellular blastoderm (stage 5) *Drosophila* embryo (B). In both, expression then fades from the anterior-most region of the embryo, as seen in the early germband extending *Tribolium* embryo (arrow in C) and the early germband extending (stage 7) *Drosophila* embryo (D); this region appears to correspond to the prospective stomodeum in both insects. Expression in *Tribolium* is further reduced as the germband extends (E), with no expression visible in the forming labrum. Transcripts are seen in a pair of lateral domains linked posteriorly by a thin line of cells. This resembles what is seen in the germband extending (stage 9) *Drosophila* embryo (F); lateral expression is joined by a thin line of cells posterior to the forming stomodeum. At this point, however, expression is still seen dorsal to the forming stomodeum in the prospective *Drosophila* labrum (expression in F is continuous dorsally in the *Drosophila* embryo, but appears broken as the procephalic lobe was split to allow the embryo to be flattened; marked by asterisks). By late germband extension in *Tribolium*, transcripts are seen posterior and lateral to the stomodeum, and in the procephalon (G). This resembles the germband extended (stage 11) *Drosophila* embryo (H); expression lies along the posterior limit of the stomodeum, and extends laterally into the procephalon. By this stage expression is no longer seen in the dorsal region corresponding to the labrum. It is also noteworthy that expression is not very extensive in the primordium of the hypopharyhngeal lobes, with transcripts only seen along its anterior extent with a thin posterior projection. hl, primordium of the hypopharyngeal lobes; lb, labrum; st, stomodeum.

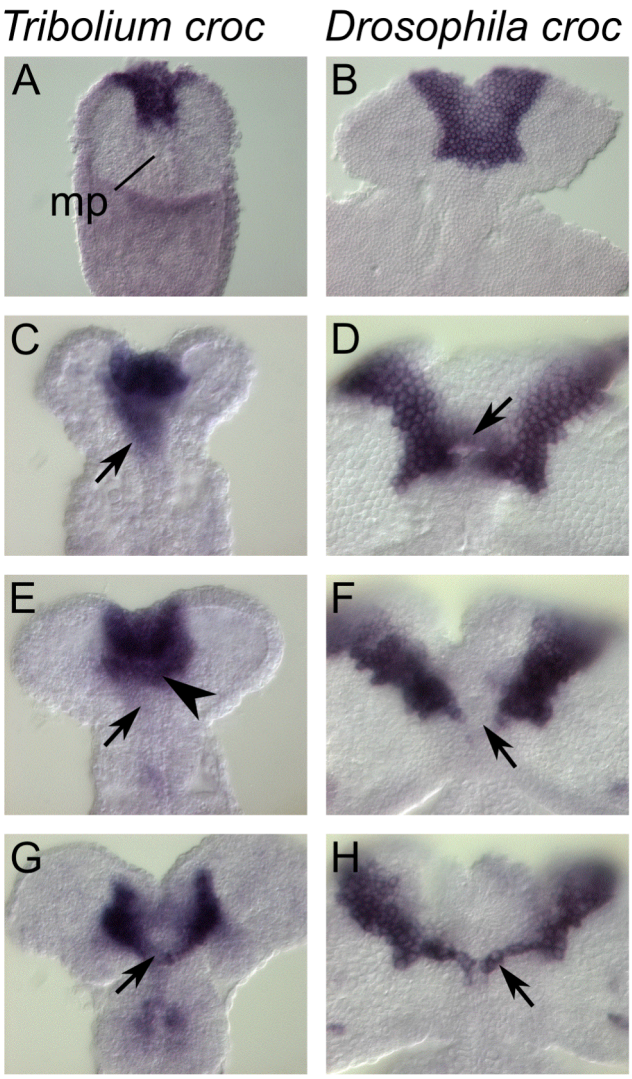


Figure 4.5. Differences in the early expression of *crocodile* orthologues in *Tribolium* and *Drosophila*. (A, C, E, G) *Tribolium*, (B, D, F, H) *Drosophila*. Ventral view of embryos, oriented with anterior up. Nomarski images. In both *Tribolium* and *Drosophila*, expression is first detected at the anterior of the early embryo, as can be seen in the gastrulating *Tribolium* embryo (A) and the cellular blastoderm (stage 5) *Drosophila* embryo (B). In *Tribolium* this expression extends posteriorly along the prospective mesoderm in the middle plate (mp in A). This expression in the *Tribolium* middle plate persists through gastrulation and is seen in the developing mesoderm of *Tribolium* (arrow in C). Similarly, expression can be seen in the mesoderm of an early germband extending (stage 7) *Drosophila* embryo (arrow in D). In both *Tribolium* and *Drosophila* this mesodermal expression fades as seen in the germband extending *Tribolium* embryo (E) and the germband extending (stage 8) *Drosophila* embryo (F). Arrows in E and F mark the region from which expression has faded. In *Drosophila* this leaves a ventral break in *croc* expression. No such break is seen in *Tribolium croc* expression (arrowhead in E). Later in germband extension, expression patterns of *Tribolium* and *Drosophila* come to resemble each other once again, as seen in the late germband extending *Tribolium* embryo (G) and the late germband extending (stage 9) *Drosophila* embryo (H). In both there are the lateral domains of expression joined ventrally by a thin domain of expression (arrows in G and H).

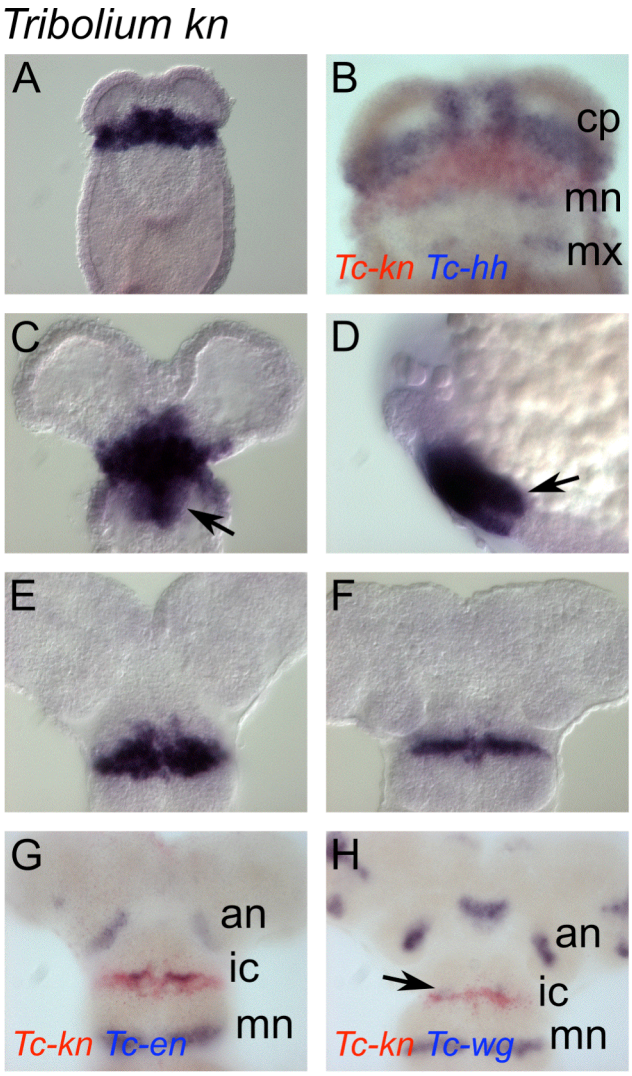


Figure 4.6. Expression of *Tribolium knot*. (A-C, E-H) Ventral view of embryos, oriented with anterior up. (D) Lateral view of embryo, oriented with anterior left. (A, C-F) Nomarski images. (B) Expression of *Tc-kn* (red) and *Tc-hh* (blue). (G) Expression of *Tc-kn* (red) and *Tc-en* (blue). (H) Expression of *Tc-kn* (red) and *Tc-wg* (blue). Brightfield images. Expression is seen in the germ rudiment (A), as a band across the embryo coincident with the posterior of the head lobes. An approximately similar staged embryo (B) shows that this band of *Tc-kn* abuts the large cephalic domain of *Tc-hh* expression anteriorly (the posterior extent of this domain appears to correspond to *Tc-hh* expression in the antennal segment by comparison to Farzana and Brown (2008), and the mandibular stripe of *Tc-hh* posteriorly. Early in germband extension (C, D), expression fades from the posterior of this domain. The persisting posterior expression (arrow) appears to be mesodermal, as it lies medially (as seen in C) and at a deeper layer (as seen in D). Later in germband extension (E) expression is lost from the posterior mesoderm and from the anterior of the domain. In a slightly older germband extending embryo, when the appendages are beginning to form (F), expression lies at the boundary of the intercalary and mandibular segments. The anterior boundary of this expression is parasegmental (G, H), lying coincident with the intercalary *Tc-en* stripes (G), but immediately posterior to the faint intercalary *Tc-wg* spots (H; arrow marks the position of *Tc-wg* spots). an, antennal; cp, cephalic; ic, intercalary; mn, mandibular; mx, maxillary.

Expression is then lost from the anterior and posterior-most parts of this early domain (figure 4.6 C-E), leaving a band of expression at the intercalary-mandibular boundary (figure 4.6 F), with an anterior coincident with the *Tc-en* expressing cells of the intercalary segment (figure 4.6 G, H).

The main difference in expression regards the mesoderm. In the early *Tribolium* embryo, there is expression across the middle plate (figure 4.6 A), which persists through germband extension (figure 4.6 C, D). In contrast there is no expression in the prospective mesoderm in the *Drosophila* blastoderm (Crozatier, *et al.*, 1996). Mesodermal *kn* is only seen later in *Drosophila* development (Seecoomar, *et al.*, 2000).

4.4.4 Expression of orthologues of cap'n'collar *and* crocodile *relative to* labial *in* Tribolium

The many striking similarities in the expression patterns between *Tribolium* and *Drosophila* for the orthologues of the three genes *cnc*, *croc* and *kn* suggest that there is expression in homologous structures. It was important, therefore, to establish whether any of the genes are expressed in the *Tribolium* intercalary segment. So far, I have only shown a clear intercalary aspect to *Tc-kn*, where its expression is coincident with the intercalary *Tc-en* stripes (figure 4.6 G), as seen in *Drosophila*. To investigate whether *Tc-cnc* or *Tc-croc* have any expression in the intercalary segment, I examined their expression relative to *Tc-lab*, which unambiguously marks the *Tribolium* intercalary segment. Neither *Tc-cnc* nor *Tc-croc* shows any expression in the *Tc-lab* domain. The "collar" of *Tc-cnc* expression lies posterior to the domain of *Tc-lab* throughout embryogenesis, showing the anterior boundary of this domain to be mandibular and the domain behind the stomodeum lies anterior to *Tc-lab* expression (figure 4.7 A and B). Similarly, the domain of *Tc-croc* expression lies anterior to the *Tc-lab* domain throughout embryogenesis (figure 4.7 C and D). Therefore neither *Tc-cnc* nor *Tc-croc* is expressed in the *Tribolium* intercalary segment.

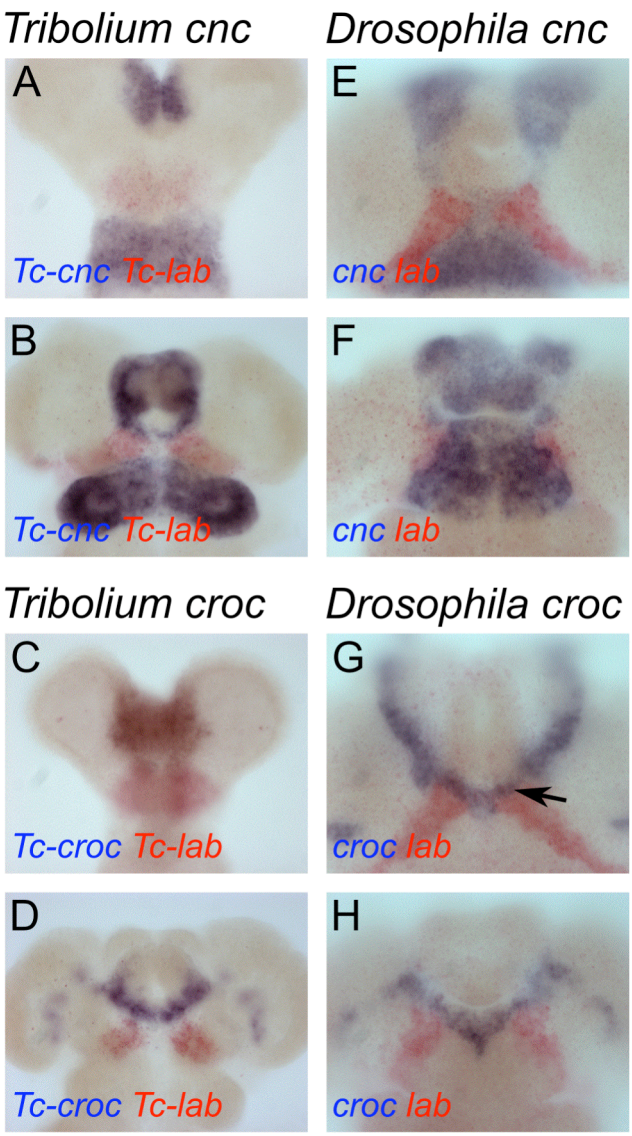


Figure 4.7. Expression of *crocodile* and *cap'n'collar* orthologues relative to *labial* orthologues in *Tribolium* and *Drosophila*. (A-D) *Tribolium* (E-H) *Drosophila*. Ventral view of embryos, oriented with anterior up. (A, B, E, F) Expression of *cnc* orthologues (blue) and *lab* orthologues (red). (C, D, G, H) Expression of *croc* orthologues (blue) and *lab* orthologues (red). Brightfield images. Expression of *Tc-cnc* does not overlap with *Tc-lab*. In the germband extending embryo (A) *Tc-cnc* lies posterior to *Tc-lab* expression. This relative expression is maintained in the germband extended embryo (B). By this stage the stomodeal domain of *Tc-cnc* is also present and lies anterior to *Tc-lab* expression. Similarly *Tc-croc* expression does not overlap with *Tc-lab*. Early in germband extension *Tc-croc* lies anterior to *Tc-lab* (C). This relative expression is maintained in the germband extended embryo (D) where *Tc-croc* is seen to lie anterior to *Tc-lab*. In the *Drosophila* germband extending (stage 9) embryo (E) *lab* expression is seen in two domains which are split across *cnc* expression. There does not appear to be any overlap between *cnc* and *lab* expression, although the stomodeal domain of *cnc* expression is faint and it is hard to make out whether it overlaps *lab* expression. This relative expression *cnc* and *lab* is maintained in the germband extended (stage 11) embryo (F). By this stage, the domains of *lab* expression are broadly separated, and are split across the domain of *cnc* expression which marks out the hypopharyngeal lobes. There is no obvious overlap between *cnc* and *lab* expression. Similarly, in the germband extending (stage 9) *Drosophila* embryo (G), the domains of *lab* expression appear to be split across the domain of *croc* expression, although there does appear to be a small degree of overlap between the expression domains (arrow). This relative expression *croc* and *lab* is maintained in the germband extended (stage 11) embryo (H), although the overlap between *croc* and *lab* expression can no longer be seen.

4.4.5 Relative expression of cap'n'collar*,* crocodile*,* knot *and* labial *in* Drosophila

The exclusion of *Tc-cnc* and *Tc-croc* from the *Tribolium* intercalary segment coupled with the similarities in expression with their *Drosophila* orthologues, strongly suggest that their expression in the primordium of the *Drosophila* hypopharyngeal lobes does not indicate that these are part of the intercalary segment either. To investigate this possibility further, and to gain a better understanding of which segments the domains of gene expression belong to, I carried out a detailed investigation of the relative gene expression patterns in *Drosophila*, to see which features of *Tribolium* expression are conserved.

In *Tribolium*, *Tc-cnc* and *Tc-croc* are expressed immediately posterior to the stomodeum, in what appears to be an overlapping domain, bounded posteriorly by intercalary *Tc-lab* expression. In *Drosophila*, when the stomodeal domain of *cnc* expression can be seen distinct from the early "collar" of expression at stage 9, the domain strongly resembles *croc* expression behind the stomodeum at the same stage (compare figure 4.2 D with figure 4.4 F). Double *in situ* hybridisation for *cnc* and *croc* in a stage 9 embryo show that these two genes are precisely co-expressed posterior to the *Drosophila* stomodeum (figure 4.8 A). The domain of *Tc-cnc* and *Tc-croc* co-expression posterior to the *Tribolium* stomodeum and anterior to the intercalary segment appears conserved in *Drosophila*.

The major difference between *Drosophila* and *Tribolium* relates to *cnc* expression: whilst the anterior stomodeal and the "collar" of expression lie adjacent to each other in *Drosophila* (figure 4.2 D, F), the stomodeal expression is separated from what is clearly a mandibular "collar" of expression in *Tribolium* (figure 4.2 E). As it has been argued that *kn* lies on the intercalary-mandibular boundary in *Drosophila* (Crozatier*, et al.*, 1999), and I have shown that this expression is conserved in *Tribolium* (figure 4.6 G and H), I looked at the expression of *croc*, relative to *kn* in Drosophila. As expected, in a stage 9 embryo, *croc* lies immediately anterior to *kn* expression but does not overlap (figure 4.8 B). This non-overlapping expression appears to be maintained in later stages as the posterior-most limit of *croc* expression moves further posterior (figure 4.8 C, D).

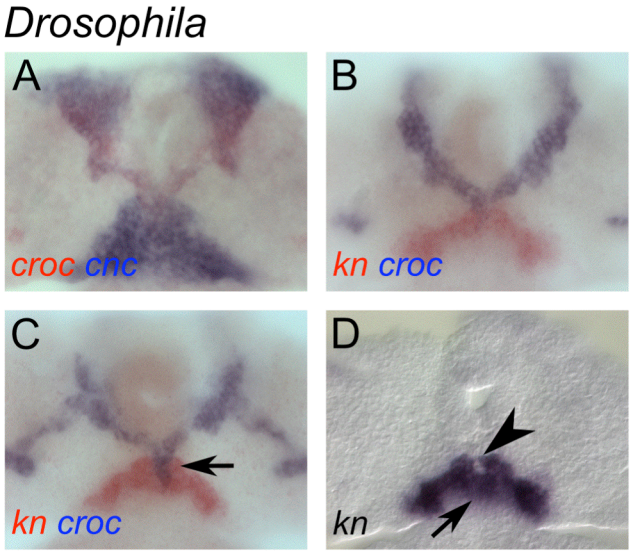


Figure 4.8. Relative expression of *crocodile*, *cap'n'collar* and *knot* in *Drosophila*. Ventral view of flattened embryos, oriented with anterior up. (A) Expression of *cnc* (blue) and *croc* (red). (B, C) Expression of *croc* (blue) and *kn* (red). Brightfield images. (D) Expression of *kn*. Nomarski image. In the germband extending embryo (stage 9), *cnc* and *croc* clearly overlap in the region posterior to the stomodeum (A). At the same stage, *kn* expression lies anterior to and abuts *croc* expression (B). In a slightly older embryo (stage 10) *croc* expression appears to have extended posteriorly through the domain of *kn* expression (arrow in C). However, closer examination of *kn* expression at the same stage (D) suggests that the midline expression of *kn* is largely mesodermal (arrow) corresponding to the late mesodermal domain of expression reported by Seecoomar *et al*. (2000). Additionally, ectodermal *kn* expression is broken at the midline (arrowhead). This suggesting that there may not actually be any co-expression of *kn* and *croc*.

By stage 11 the expression of *kn* has largely faded (Crozatier, *et al.*, 1999) so this expression could not be followed further.

My results suggest that *Drosophila* and *Tribolium* differ in the relative position of their stomodeal *cnc* and *croc* expressing domain: in *Tribolium* this domain is separated from the mandibular expression of *Tc-cnc* and *Tc-kn* by *Tc-lab*, whilst in *Drosophila* the stomodeal *cnc* and *croc* expression lies adjacent to the *cnc* and *kn* expression. *Tc-cnc* and *Tc-croc* do not show any overlap with the intercalary marker *Tc-lab* in *Tribolium*. I therefore asked whether this situation was conserved in *Drosophila*. Double *in situ* hybridisation for *cnc* and *lab* and *croc* and *lab* shows that whilst at stage 9 there is some possible overlap of expression between *cnc* and *croc* with *lab* expression (figure 4.7 E and G), by stage 11 there is no overlap between *cnc* or *croc* and *lab* (figure 4.7 F and H). *lab* expression appears to be split by the domain of stomodeal *cnc* and *croc* expression.

4.5 Discussion

The results presented here show that the three genes with a role in the development of the *Drosophila* hypopharyngeal lobes (*cnc*, *croc* and *kn*) have multiple conserved features of expression in *Tribolium*. However, comparison with *Tc-lab* expression, which unambiguously marks the intercalary segment in *Tribolium*, demonstrates that *Tc-croc* and *Tc-cnc* are not expressed in this segment in the beetle. Only *Tc-kn* has an obviously intercalary aspect to its expression. I further demonstrated that the differences between *Drosophila* and *Tribolium* can be explained by the movement of a single domain of expression. Both insects have a region behind the stomodeum which expresses *cnc* and *croc* orthologues. In *Tribolium*, this domain is separated from the more posterior expression of *cnc* and *kn* orthologues by *Tc-lab* expression, whilst in *Drosophila*, these two domains are adjacent, splitting the expression of *lab*.

4.5.1 A derived topology for the Drosophila *embryonic head*

In the light of these results I propose that the differences in expression patterns can be explained by a simple difference in the topology of the embryo, in the context of conserved expression in homologous structures. Both *Drosophila* and *Tribolium* share a segmental register of gene expression, with *cnc* orthologues in the mandibular segment, *lab* orthologues in the intercalary segment and *kn* orthologues along the boundary. There is also conserved expression of *cnc* and *croc* orthologues associated with the stomodeum. Where they differ is in the position of the stomodeum (and the associated expression of *cnc* and *croc* orthologues) in the segmental register: in *Tribolium*, as in other insects, the stomodeum lies anterior to the intercalary segment (and *Tc-lab* expression), whilst in *Drosophila* it has a more posterior position, lying immediately anterior to the mandibular segment (splitting the *lab* expression). These relative patterns of gene expression are summarised in figure 4.9. *Drosophila* differs

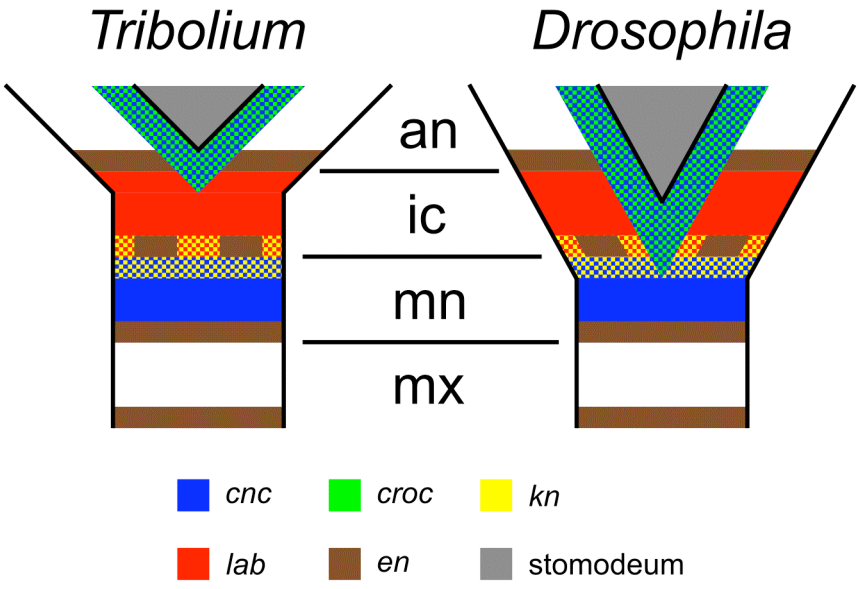


Figure 4.9. Relative expression patterns of *cap'n'collar*, *crocodile*, *knot* and *labial* orthologues in *Tribolium* and *Drosophila*. Schematic showing the relative expression patterns of *cnc* (blue), *croc* (green), *kn* (yellow) and *lab* (red) orthologues relative to the stomodeum (grey) in *Tribolium* and *Drosophila*. The positions of the antennal, intercalary, mandibular and maxillary segments are marked by their *en* expression (brown). *Tribolium* and *Drosophila* show the same patterns of gene expression, except that in *Drosophila*, the stomodeum and its associated expression has a more posterior position in the segmental register, lying anterior to the mandibular segment and *cnc* expression, while splitting the intercalary segment and *lab* expression. Several aspects of relative gene expression from Crozatier *et al.* (1996), Mahaffey *et al.* (1989), Mohler *et al.* (1995) and Nie *et al.* (2001) are included in the schematics. an, antennal; ic, intercalary; mn, mandibular; mx, maxillary.

from *Tribolium* in that the intercalary segment is split by the stomodeum. As the hypopharyngeal lobes of *Drosophila* derive from *cnc* and *croc* expressing cells they do not belong to the intercalary segment, as suggested previously (Mohler, *et al.*, 1995, Seecoomar, *et al.*, 2000). Rather, the lobes appear to be a composite structure; they are largely mandibular in origin, deriving from the "collar" of *cnc* expressing cells, whilst the anterior-most portion derives from the *cnc* and *croc* expressing cells associated with the stomodeum.

This difference in topology is likely to be related to the derived mode of head embryogenesis seen in *Drosophila*. As has already been seen, the embryonic head of *Drosophila* has the derived appearance of a series of lobes. As part of this restructuring of the embryonic head, it seems that the intercalary segment has come to lie dorsal to the mandibular segment rather than the more anterior position in other insects. It has previously been noted that the segmental axis of *Drosophila* has a marked S-shaped deflection (Schmidt-Ott and Technau, 1992). Consequently, the stomodeum now lies in front of the mandibular segment. It is important to remember that the mouth of an arthropod is ancestrally an anterior structure. In several outgroup taxa to the arthropods such as the tardigrades as well as in various stem arthropods such as *Kerygmachela*, it has a terminal position, and has subsequently been ventralised in the arthropods (Budd, 2001). This means that historically, the mouth, and therefore the stomodeum do not belong to any particular segment. It is therefore, not unreasonable to argue for a movement in its position in the segmental register in association with a dramatic change in early embryonic movements.

4.5.2 Derived features of labial *expression in* Drosophila

This interpretation is in marked contrast Rogers and Kaufman's (1997) proposal that the *Drosophila* hypopharyngeal lobes are the ventral part of the intercalary segment, which lies behind the stomodeum as in other insects (as summarised in figure 4.1 B). Their argument was based on the observation that the hypopharyngeal lobes derive from tissue that previously appeared to express the intercalary segment marker *lab*. In their description of *lab* expression, Diederich *et al.* (1989) show that *lab* is expressed across

the embryo. Moreover, my double *in situ* hybridisations give some support for the co-expression of the stomodeal domain of *cnc* and *croc* with this *lab* domain in the earlier embryonic stages.

However, I do not believe that this contradicts the interpretation of the *Drosophila* intercalary segment I have presented. My arguments for homology are based on shared details of *cnc* and *croc* expression in *Drosophila* and *Tribolium*. It seems very unlikely that any new expression domain in the ventral intercalary segment in flies would resemble so strongly the expression and modulations of the domain anterior to the intercalary segment in beetles. Whilst it may be true that these cells expressing *cnc* and *croc* do transiently express *lab* earlier in embryogenesis, this difference in expression with *Tribolium* is most likely a result of the derived embryogenesis of *Drosophila*. The blastoderm of *Drosophila* is topologically a very different environment to the germ rudiment of *Tribolium* and it is therefore likely that the early regulation of gene expression differs between the two. It is possible that *lab* is first expressed in a more extensive domain in the fly which is subsequently refined. Homology should not be assigned on the basis of shared expression of a single gene. Rather, detailed similarities in the relative positions and timings of expression for several genes, such as those presented here should be used to assign homology.

4.5.3 Differences in the early embryology of Drosophila *and* Tribolium

This argument relies on the homology of the expression of *cnc* and *croc* orthologues behind the stomodeum of *Drosophila* and *Tribolium*. However, there are differences in the early modulations of this domain, in particular for *croc*. In *Drosophila* a gap was reported in the ventral expression of *croc* which was bridged later by the stomodeal domain. Such a gap in expression could not be seen in the *Tribolium* orthologue suggesting that the stomodeal expression may not be a domain which appears late in embryogenesis, but instead part of the early anterior domain of expression. It is therefore necessary to address what lies behind these early differences in expression.

The position of the foregut anlage, from which the stomodeum forms, differs between *Drosophila* and *Tribolium* (see figure 4.10). In *Drosophila*, the ventral furrow (from which the prospective mesoderm forms) stops posterior to the foregut anlage (de Velasco, *et al*., 2006). In contrast, the prospective mesoderm of *Tribolium* (the middle plate) runs to the anterior-most point of the embryo (Handel, *et al*., 2000). The precise position of foregut anlage has not yet been fate mapped in the beetle, but as it is an ectodermal structure it must be bisected by the prospective mesoderm.

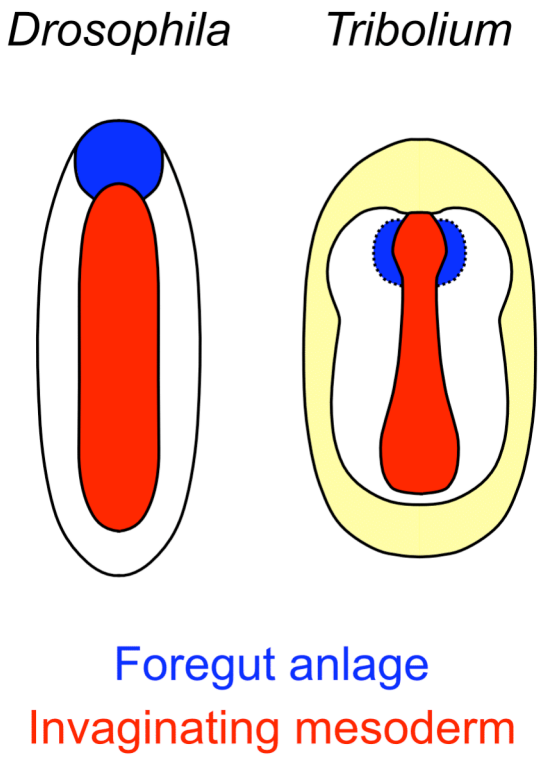


Figure 4.10. Differences in the location of the foregut anlage in *Drosophila* and *Tribolium*. Schematics showing the position of the foregut anlage (blue) relative to the invaginating mesoderm (red) in a late blastodermal *Drosophila* embryo and a germ rudiment *Tribolium* embryo (yellow represents yolk). Embryos shown in ventral view with anterior up. In *Drosophila* the foregut anlage lies anterior to the site of meodermal invagination (the ventral furrow). In contrast, the prospective mesoderm in *Tribolium* (the middle plate) runs to the anterior point of the embryo and therefore splits the foregut anlage (an ectodermal structure). The dashed lines surrounding the *Tribolium* foregut anlagen indicate that its position is predicted, as it has not been fate mapped. Schematics based on de Velasco *et al*. (2006) for the position of the *Drosophila* mesoderm relative to the foregut anlage and Handel *et al*. (2005) for the location of the prospective mesoderm of *Tribolium*.

These differences in the relative positions of the prospective mesoderm and the foregut anlage have clearly altered the embryology of the prospective stomodeum. In *Drosophila*, the foregut anlage is always a single domain, whilst in *Tribolium* the two halves of the foregut anlage must come together as the prospective mesoderm invaginates. Given this difference in early embryology, it is not surprising that the early gene expression associated with the forming stomodeum differs. Interestingly, the position of the forming mesoderm posterior to the foregut anlage in *Drosophila* means that the developing mesoderm undergoes anterior migration from the anterior ventral furrow (de Velasco, *et al*., 2006); such migrations do not occur in other *Tribolium* where the prospective mesoderm runs to the anterior of the embryo. It seems that these differences in mesodermal embryology also appear to be associated with differences in gene expression; as I showed, *Tc-kn* is expressed in the head mesoderm from early on in *Tribolium*, but only at a later stage in its *Drosophila* orthologue. The expression patterns must be interpreted in the context of the embryology.

4.5.4 Implications for the Drosophila *head fate map*

My results are in agreement with the arguments made by Mohler *et al*. (1995) and Seecoomar *et al*. (2000) that the hypopharyngeal lobes do not represent a major embryonic component of the intercalary segment. However, there are important differences from my interpretation. Whilst both these studies argue that the hypopharyngeal lobes belong to the mandibular segment, I have shown that some aspects of gene expression that have previously been attributed to the hypopharyngeal lobes (namely *croc* expression and part of *cnc* expression) do not belong to the mandibular segment. Rather I argue that they belong to a distinct domain associated with the stomodeum. Therefore the lobes are a composite structure, the posterior being part of the mandibular segment and the anterior deriving from cells associated with the stomodeum.

This has implications for the *Drosophila* head fate map and in particular the primordium of the posterior pharyngeal wall (ppw). Jurgens *et al*. (1986) showed that ablation of the hypopharyngeal lobes led to the loss of the ppw. However, Mohler *et al*.

(1995) questioned this interpretation, arguing that cells originating at the base of the labrum were found in the ppw. They suggested that some of these cells may have been ablated by Jurgens *et al.* (1986) as well as the cells residing in the hypopharyngeal lobes. My results support the view of Jurgens *et al.* (1986). Mutants of *cnc* and *croc* lose the ppw and I have shown that these genes are co-expressed behind the stomodeum in part of the hypopharyngeal lobes. Therefore, it seems likely that they are involved in the differentiation of the hypopharyngeal lobes to a pharyngeal fate. This is not to say that cells at the base of the labrum do not also contribute to the ppw. The *cnc* and *croc* expressing domain and the base of the labrum lie immediately posterior and anterior to the stomodeum respectively. Given that the ppw lies immediately anterior to the oesophagus, an origin from cells immediately posterior and anterior to the stomodeum would be expected.

Functional work in *Tribolium* would be required to confirm whether this domain expressing *cnc* and *croc* gives rise to part of the foregut. Interestingly Rogers *et al.* (2002) also identified a similar domain of *cnc* expression in the milkweed bug *Oncopeltus fasciatus*. Whilst they argued that it belonged to the anterior intercalary segment, this assignment was made in the absence of any markers. In the light of my results it seems likely that this domain is homologous to the stomodeal domain that I have identified in *Drosophila* and *Tribolium*. It therefore seems that this expression domain is conserved more widely in the insects, although Rogers *et al.* (2002) do not report the presence of this domain in the firebrat, *Thermobia domestica*.

4.6 Conclusions

I addressed the issue of what constitutes the intercalary segment in the model organism *Drosophila melanogaster*, specifically asking whether a pair of lobes behind the stomodeum – the hypopharyngeal lobes – constitute the ventral part of the intercalary segment. I took a comparative approach and demonstrated that the genes expressed in

the *Drosophila* hypopharyngeal lobes are expressed in homologous structures in the red flour beetle *Tribolium castaneum*, but that these genes are not expressed in the intercalary segment of either insect. On this basis, two of the genes previously implicated in patterning the *Drosophila* intercalary segment – *cnc* and *croc* – do not appear to have a role in the development of the segment. Therefore, very few genes are known with a clear role in patterning the intercalary segment in *Drosophila* and other insects; only *kn* and *lab* have conserved expression patterns between the beetle and fly. In the following chapter I address this issue, using the comparative approach presented here to find more candidate genes for patterning the insect intercalary segment.

Chapter 5:
The development of the intercalary segment and the search for new genes

5.1 Summary

Little is known about how the intercalary segment develops in the model organism *Drosophila melanogaster* or any other insect. In this chapter I attempt to further what is known about insect intercalary segment development. I present a screen to identify additional candidate genes for patterning the segment, searching for genes with conserved expression patterns in the intercalary segments of *Drosophila* and *Tribolium castaneum*. I first identified genes with expression in the intercalary segment of *Drosophila*, by searching through the Berkley *Drosophila* Genome Project expression pattern database. I then identified orthologues of these genes in *Tribolium* genome using the BeetleBase database. I finally carried out *in situ* hybridisations for these genes in *Tribolium* to see if the intercalary segment expression pattern in *Drosophila* is conserved. Using this screen I identified four genes with expression patterns associated with the intercalary segment; one with expression in the posterior intercalary segment ectoderm and three with expression in the intercalary segment mesoderm. Moreover, I suggest that the three genes with conserved expression in the intercalary segment mesoderm are specifically expressed in developing hemocytes; possibly the major mesodermal derivative of the segment.

5.2 Introduction

Having resolved what constitutes the intercalary segment of the model organism *Drosophila melanogaster*, it seems that very few genes appear to have a clear role in patterning the intercalary segment, even in the fly. As we have seen, the head gap-like genes have a well-characterised role in establishing the segment, but this is not conserved in red flour beetle *Tribolium castaneum*. Only *knot* (*kn*) has a definite role in the development of the segment in *Drosophila* and a conserved expression pattern in *Tribolium*. *cap'n'collar* (*cnc*) and *crocodile* (*croc*) which had previously been implicated in patterning the segment in the fly, are not expressed in the intercalary segment in either *Drosophila* or *Tribolium*. Also, whilst the *Drosophila* expression pattern of *labial* (*lab*) does now appear to be the same as in other insects, *lab* mutants in the fly and RNAi knock downs in the milkweed bug *Oncopeltus fasciatus* have shown no obvious phenotype relating to the intercalary segment.

Given that the intercalary segment has a very derived morphology in comparison to the ancestral crustacean second antennal segment, there is clearly a lot about its development that is not known. It is probable that there are still a number of genes that play a role in the patterning of the segment that have not yet been discovered. For example, the segment possesses no appendages and no genes have been implicated in any developmental basis to this. Also, the structure of the mesodermal somites is unlike that of any other segment, yet it is not known how they differentiate differently to any other segment.

Fortunately, studies of development in the arthropods and in particular the insects are greatly aided by the many resources created for the study of *Drosophila*. *Drosophila* is arguably the best-studied organism in terms of its developmental genetics, with only *Caenorhabditis elegans* being understood to anything like a comparable level. Consequently, the techniques available for studying the fly surpass those available for any other arthropod. As well as a range of embryological tools for studying *Drosophila*, there are an increasingly large number of online resources. Many of these are provided by the Berkeley *Drosophila* Genome Project (BDGP;

http://www.fruitfly.org/). The BDGP is a consortium of the *Drosophila* Genome Centre, whose goals are to finish the sequence of the euchromatic genome of *Drosophila melanogaster* to high quality and to maintain biological annotations of this sequence. One of the further aims of the BDGP is to characterise the sequence and expression of *Drosophila* cDNAs. As part of this project they have used high-throughput RNA *in situ* hybridisation to establish a database of gene expression patterns during embryonic development of *Drosophila* (Tomancak, *et al.*, 2002).

This BDGP expression pattern database provides a source of expression patterns for a number of genes, including many as yet unstudied genes, only known by their annotation identifier (CGnnnn or CGnnnnn). The expression patterns in the database are grouped into various developmental stages, and within each stage are annotated by the embryonic structures each gene is expressed in. These annotations allow the database to be searched for expression in particular embryonic structures during the different developmental stages. A number of genes expressed in a particular part of the embryo can be recovered, several of which may be unstudied genes not previously implicated in the development of that structure. Given that lack of knowledge about the development of the intercalary segment, the database provides a potential source of genes with localised expression in and around the segment. Such genes are obvious candidates for a role in patterning the *Drosophila* intercalary segment.

However, head development in *Drosophila* has a number of derived features for an insect. As was discussed in chapter 4, several features of *Drosophila* head embryology appear derived, most likely as a result of head involution. These morphogenetic movements almost certainly have unique gene expression patterns associated with them. Also, several features of early *Drosophila* head development are not conserved in other insects. It has already been seen that the overlapping expression domains of the gap-like genes that play a role in the segmentation of the *Drosophila* head are not conserved in *Tribolium*. In addition, the morphogen Bicoid that is involved in regulating much of early gene expression in the head is unique to the higher Diptera (Lynch, *et al.*, 2006, Schröder, 2003). It is, therefore, unclear to what extent a solely *Drosophila* based model of intercalary segment development would be applicable to other insects.

In spite of these derived features of *Drosophila* head development, it is not unreasonable to assume that the conserved morphological features of the fly intercalary segment would be underpinned by at least some features of development conserved in other insects. For example, despite the many difference in early head development, the hox genes *Deformed* (*Dfd*), *Sex combs reduced* (*Scr*) and *proboscipedia* (*pb*) appear to have several conserved functions in patterning the posterior head and mouthparts in a range of insects (Hughes and Kaufman, 2002b). Searching for the features of *Drosophila* intercalary segment development that are conserved in other insects would seem to be a productive approach to studying insect intercalary segment development.

Based on this premise I have carried out a screen to identify new candidate genes for patterning the intercalary segment. Specifically, I have looked for genes with conserved expression patterns between *Drosophila* and *Tribolium*. Firstly, using the BDGP expression pattern database, I identified genes with an intercalary segment expression pattern in *Drosophila*. I identified orthologues in *Tribolium* and cloned partial cDNAs. I examined their expression patterns in *Tribolium* to see if they had conserved expression in the intercalary segment. This approach successfully recovered a set of genes with conserved expression patterns between the fly and beetle.

5.3 Materials and Methods

5.3.1 Screening the BDGP expression pattern database

The BDGP expression pattern database was searched for genes with expression patterns relating to the intercalary segment. The annotations do not go down to the level of structures as specific as individual segments. Therefore, to find expression patterns potentially relating to the intercalary segment, using the *Basic Search* option the database was searched for genes with expression in the procephalic ectodermal and head mesodermal anlagen and primodia (see table 5.1 for the precise search terms used).

Table 5.1. Search terms used in the *Basic search* of the BDGP expression pattern database. Combinations of the embryonic stage and corresponding embryonic structure are given.

Stage	Embryonic structure
4-6b	procephalic ectoderm AISN
7-8	procephalic ectoderm Anlage
9-10	procephalic ectoderm PR
4-6b	head mesoderm AISN
7-8	head mesoderm anlage
9-10	head mesoderm P2 PR

The search recovered a number of genes with a range expression patterns in the head. Many genes recovered by the search were expressed in structures that were not relevant to the intercalary segment. For example, only a subset of expression patterns in the *procephalic ectoderm Anlage* for a stage *7-8* embryo would be relevant to the intercalary segment. The expression patterns for all the genes recovered by each search were inspected by eye, and genes with expression in structures deemed to be relevant to the intercalary segment were selected (see section 5.3.2). Annotations are described by the BDGP as a "work in progress" (http://www.fruitfly.org/ex/FAQ.htm) and so there could be a level of inaccuracy. For example expression in some stages may be missed or germ layers could be misidentified. Therefore, all stages were inspected for any possible intercalary segment expression, not just the stage specified in the search. All genes with possible intercalary segment expression patterns were selected, not just regulatory genes such as transcription factors and signalling proteins, which would be expected to have a developmental role. Other genes with localised expression patterns could have roles in the differentiation of specific segmental structures.

5.3.2 Intercalary segment expression patterns

Potential expression in the intercalary segment ectoderm was based on similarity to the two genes known to have localised expression in the segment, namely *lab* and *kn* (see chapter 4). These domains are summarised in figure 5.1 A-C. Genes with expression patterns resembling these domains, or with distinctive patterns within or bordering these domains were selected.

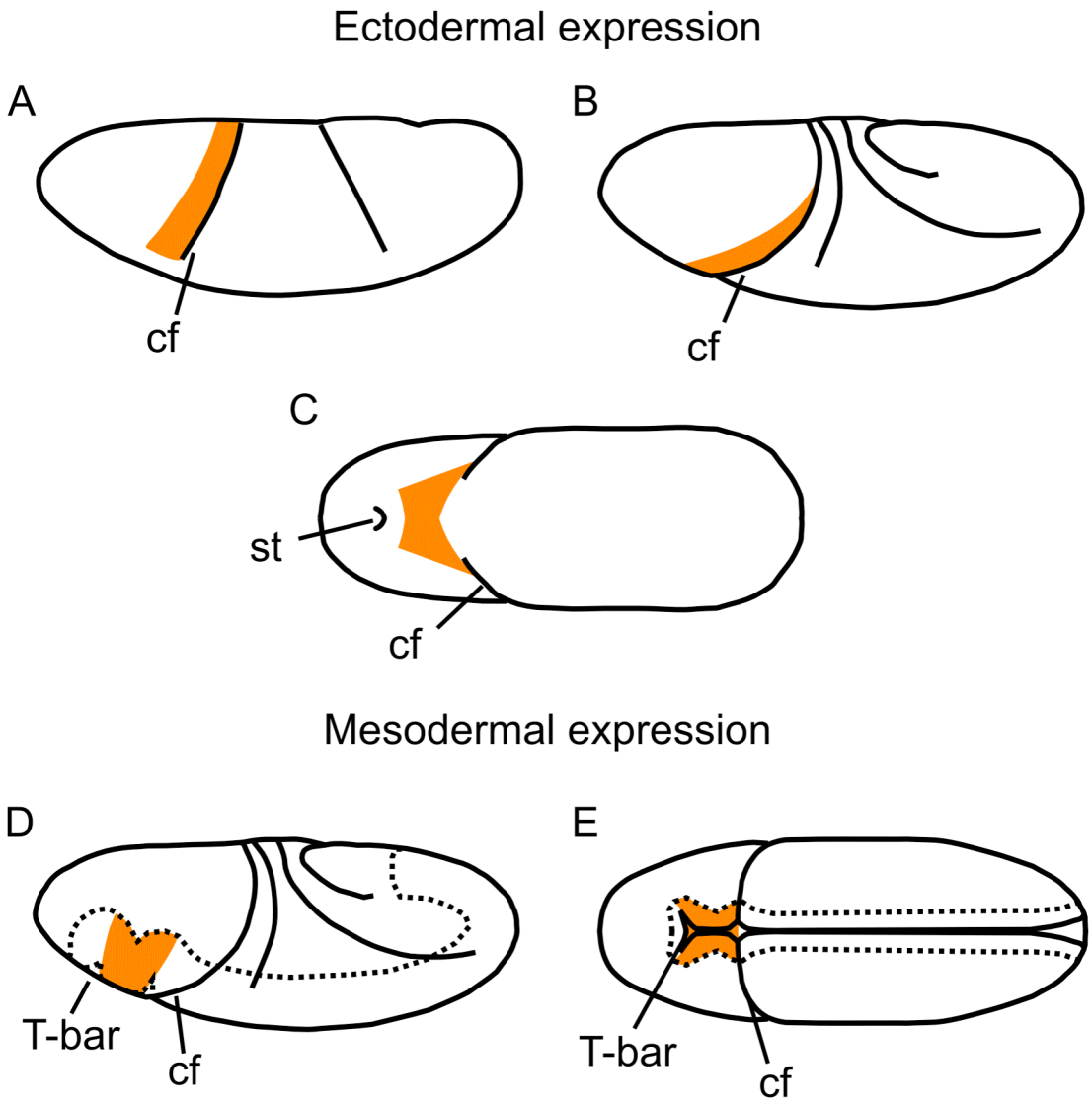


Figure 5.1. Schematics showing domains of gene expression associated with the intercalary segment. Areas of interest are shown in orange. A-C show the domain of expression in the ectoderm in a late blastoderm embryo (A), gastrulating embryo (B) and germband extending embryo (C). The position of the cephalic furrow (cf) marking the back of the procephalon is shown in A, B and C, and the position of the stomodeum (st) is shown in C. D and E show the domain of expression in the mesoderm during gastrulation. The cephalic furrow and the anterior "T-bar" of the ventral furrow (see de Velasco, *et al*., 2006) are shown in D and E. The extent of the internalised mesoderm is marked by the dashed line.

Potential expression in the intercalary mesoderm was as described by de Velasco *et al.* (2006). They argue that the region between the front of the ventral groove (the "T-Bar") and the cephalic furrow of the gastrulating embryo (their primary head mesoderm domains B and C) belongs to the intercalary segment (summarised in figure 5.1 D, E). Genes with expression in these areas also were selected. For older embryos, genes with expression patterns in the posterior head mesoderm were also selected.

5.3.3 Identification of Tribolium *orthologues*

Tribolium orthologues of the genes selected from the BDGP expression pattern database search were identified by a BLAST search of the *Tribolium* genome followed by a reciprocal BLAST search of *Drosophila* proteins, as described in section 2.3.5 (accession numbers for *Drosophila* query sequences are given in appendix 1 table A1.3). Several of the *Drosophila* candidate genes had multiple isoforms in GenBank. To identify the most appropriate isoform for use as the query sequence in the BLAST search, the different isoforms were aligned using MacClade 4.06, and the isoform which represented most sequence was chosen. In cases where different isoforms had very different sequences, all were used as query sequences.

5.3.4 Tribolium in situ *hybridisation screen*

Tribolium stocks were maintained as described in section 2.3.1, and embryos were collected and fixed as described in sections 2.3.2, 2.3.3 and 2.3.9. Partial cDNAs of the *Tribolium* orthologues of *Drosophila* intercalary segment genes were amplified by PCR from *Tribolium* cDNA (primer sequences are given in appendix 2, table A2.2) and cloned (see sections 2.3.4 and 2.3.6). *In situ* hybridisation was carried out as described in section 2.3.10 using DIG labelled probes (see section 2.3.8); 5 μl of probe was used. In the cases when no stain showed up within 4-5 hours, embryos were left to develop the stain at 4ºC overnight. To confirm that when probes did not show any localised expression it was not due to a general problem affecting the batch of embryos used or the buffers used, a positive control was run at the same time, using the same batch of

embryos and buffers. This was typically using the probes for *Tc-cnc*, *Tc-croc* or *Tc-kn* (see chapter 4). Embryos were prepared and imaged as described in section 2.3.13.

During the course of the screen it had become apparent that preabsorbing the antibody and using less probe reduced background for fluorescein labelled probes and marginally improved the signal to background ratio for DIG labelled probes (see section 2.3.12). Therefore, to produce clearer expression patterns, for these final genes 0.5 μl probe was used and the anti-DIG antibody was preabsorbed to *Tribolium* embryos. To investigate whether the altered conditions could have recovered localised expression patterns that would have been missed under the previous set of conditions, a batch of genes was chosen for which *Tribolium in situ* hybridisation under the original conditions displayed either high background or no localised expression pattern. These were repeated under the new conditions. This did not affect whether an expression pattern was recovered or not.

5.3.5 Detailed examination of Tribolium *and* Drosophila *expression patterns*

For a few of the genes, a subsequent more detailed examination of the *Tribolium* expression patterns was carried out and the *Drosophila* expression patterns were also examined. *Tribolium in situ* hybridisation was carried out using the DIG labelled probes synthesised in section 5.3.4. Other *Tribolium* techniques were as described in section 5.3.4. For *Drosophila*, complete cDNAs were ordered from the *Drosophila* Gene Collection as described in section 2.3.7 (clone names are given in appendix 3, table A3.1) and DIG labelled probes were synthesised as described in section 2.3.8. *Drosophila* stock maintenance, embryo collection and fixation were as described in sections 2.3.1, 2.3.2, 2.3.3 and 2.3.9. For both *Tribolium* and *Drosophila in situ* hybridisation was carried out as described in section 2.3.10; conditions were varied to reduce background as described in section 2.3.12. *Tribolium* and *Drosophila* embryos were prepared and imaged as described in section 2.3.13.

5.4 Results

5.4.1 BDGP expression pattern database screen

Screening the BDGP expression pattern database recovered 63 genes with expression patterns relating to the *Drosophila* intercalary segment. These expression patterns are shown in figure 5.2. Out of these 63 genes, 21 had possible expression in the intercalary ectoderm. 14 of these genes had expression patterns associated with the posterior of the procephalon (by comparison to the search images in figure 5.1 A and B). The remaining 7 showed other distinctive domains of expression associated with the intercalary ectoderm (expression within or along the edges of the areas of interest marked in figure 5.1 A-C). 42 genes had expression patterns relating to the intercalary segment mesoderm. Of these, 24 had expression patterns associated with the early intercalary segment mesoderm (by comparison to the search images in figure 5.1 D and E), whilst 18 were expressed later on in the posterior head mesoderm. The domains of expression of these genes are summarised in table 5.2. Additionally, out of the 63 genes, 33 were previously unstudied genes only known by their annotation identifier. The rest were named genes, which had been studied to different extents, but as yet had not been explicitly implicated in patterning the intercalary segment. For simplicity, genes will be referred to by their annotation identifiers even if they have been named.

5.4.2 Identifying Tribolium *orthologues*

It was not possible to identify *Tribolium* orthologues for all 63 of these *Drosophila* genes using a BLAST search of the *Tribolium* genome followed by a reciprocal BLAST search of the *Drosophila* protein database. I will now describe the different outcomes of the reciprocal BLAST search, which are summarised in table 5.3.

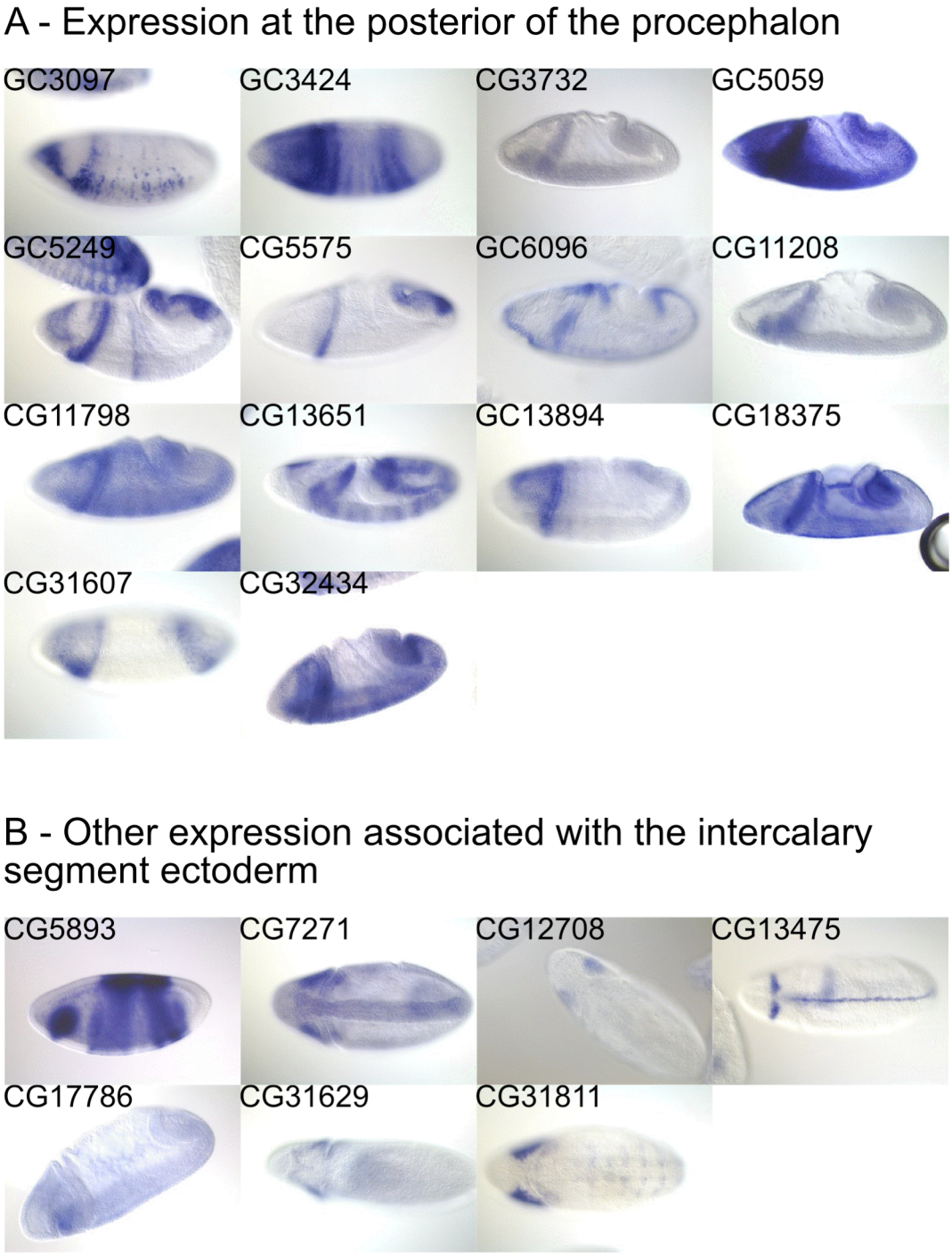
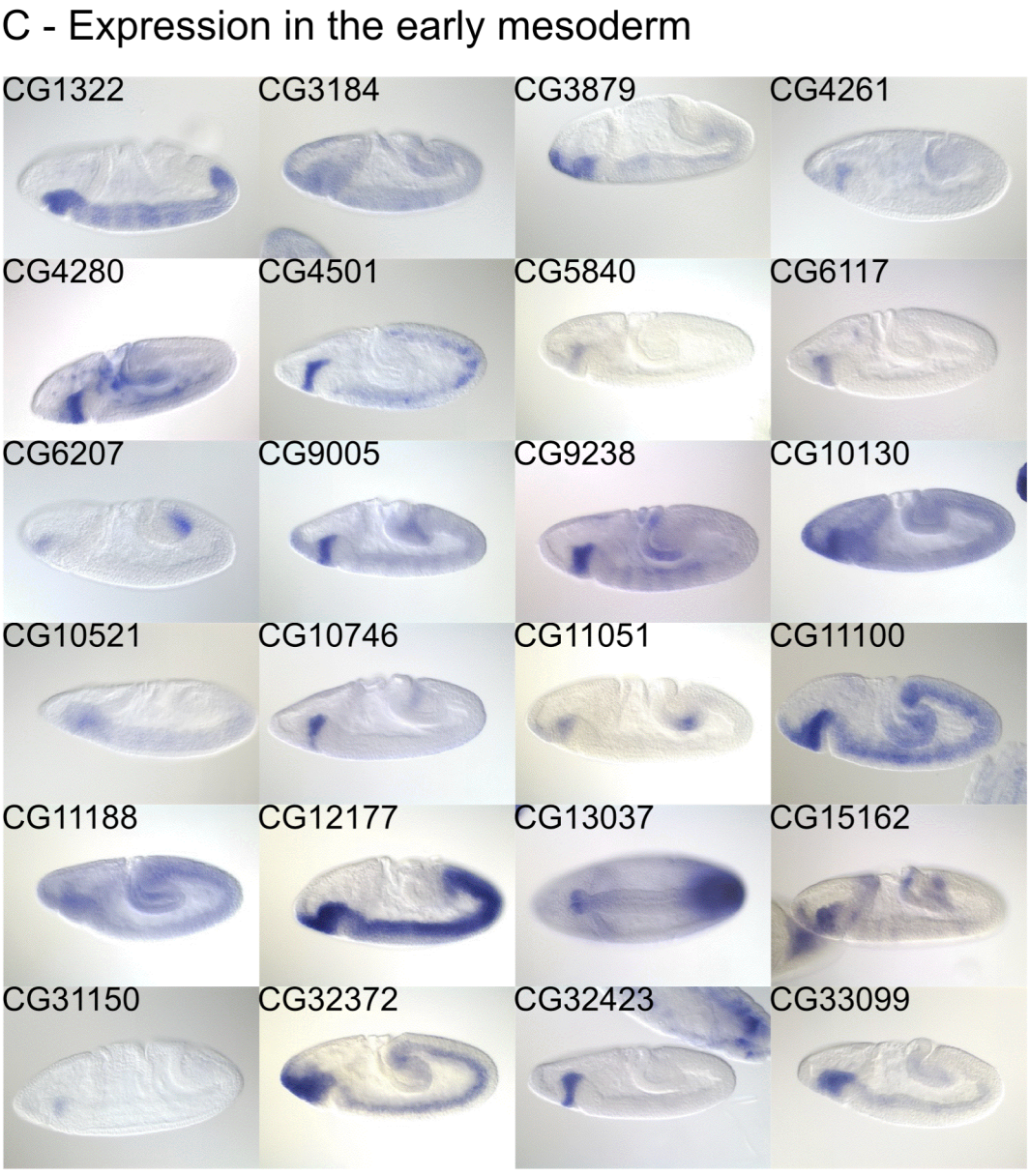
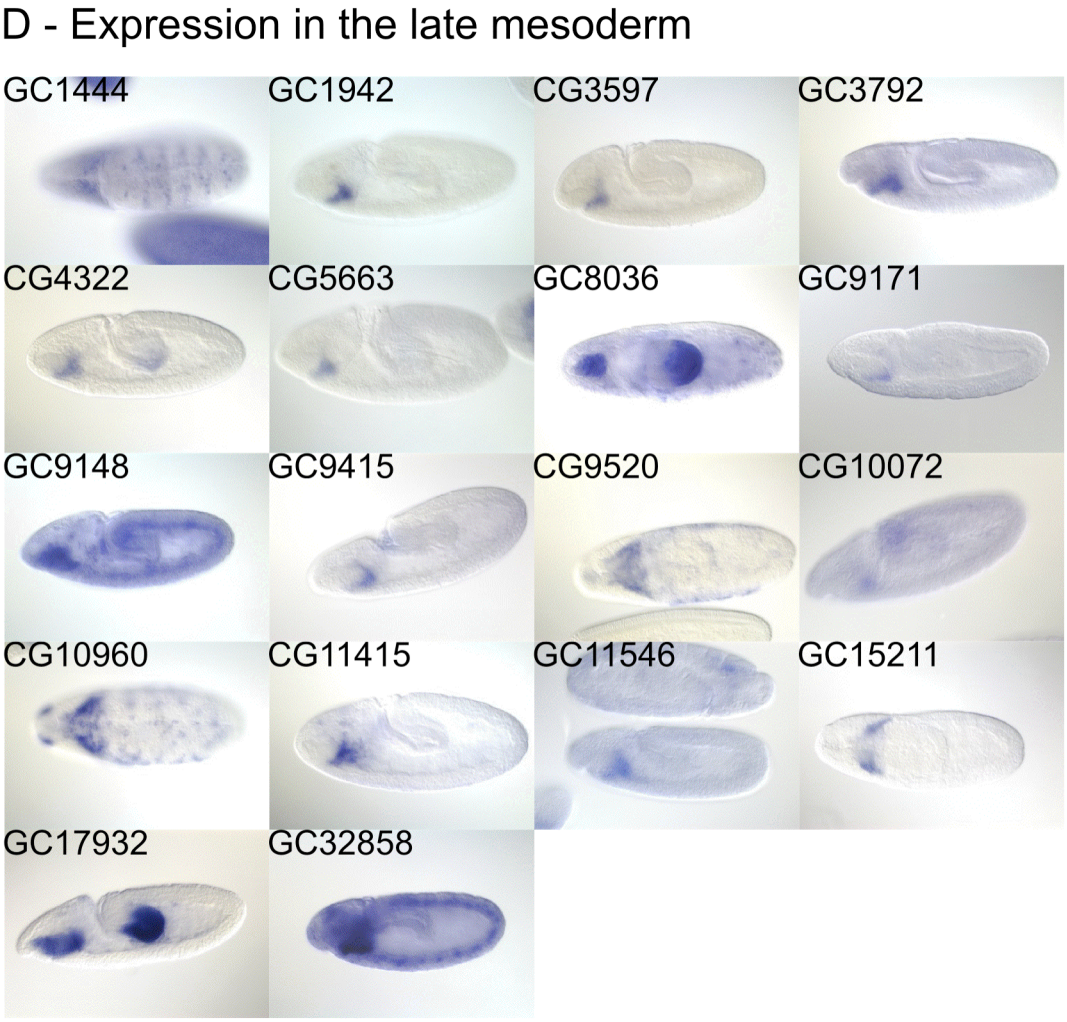


Figure 5.2. ***Drosophila*** **gene expression patterns relating to the intercalary segment.** Embryos orientated with anterior left. Most embryos are in lateral view with dorsal up. Some embryos are in ventral view. Stages vary between genes. Genes with similar expression patterns have been grouped together. Expression patterns are grouped according to whether the gene is expression in the ectoderm (A and B) or the mesoderm (C and D). Ectodermal genes with expression patterns at the posterior of the proecphalon (expression associated with the cephalic furrow) are shown in A. Other ectodermal expression patterns potentially relating to the intercalary segment are shown in B. Mesodermal genes with expression patterns in the gastrulating embryo between the T-bar and cephalic furrow (as in de Velasco, *et al*., 2006) as shown in C. Mesodermal genes with later expression in the posterior procephalon are shown in D. Genes are labelled according to their annotation identifier. Images have been taken from the BDGP expression pattern database (http://www.fruitfly.org/cgi-bin/ex/insitu.pl).



(Figure 5.2 continued)



(Figure 5.2 continued)

Table 5.2. *Drosophila* genes recovered by searching the BDGP expression pattern database for expression in the intercalary segment. Genes are grouped according to similar expression patterns (see figure 5.2). Genes are named according to their annotation identifier; where the gene has been named this is given as well. The total number of genes expressed in each region is indicated at the bottom.

Ectodemal expression		Mesodermal expression	
Posterior procephalon	Other	Early	Late
CG3097	CG5893 - *Dichaete* (*D*)	CG1322 - *Zn finger homeodomain 1* (*zfh1*)	CG1444
CG3424	CG7271	CG3184	CG1942
CG3732	CG12708	CG3879 - *Multi drug resistance 49* (*Mdr49*)	CG3597
CG5059	CG13475 - *HGTX*	CG4261 - *Helicase 89B* (*Hel89B*)	CG3762 - *Vha68-2*
CG5249 - *Blimp-1*	CG17786	CG4280 - *croquemort* (*crq*)	CG4322 - *moody*
CG5575 - *ken and barbie* (*ken*)	CG31629	CG4501 - *bubblegum* (*bgm*)	CG5663
CG6096 - *E(spl) region transcript M5* (*HLHm5*)	CG31811 - *centaurin gamma 1A* (*cenG1A*)	CG5840	CG8036 - *Dipeptidase C* (*Dip-C*)
CG11208		CG6117 - *cAMP-dependent protein kinase 3* (*Pka-C3*)	CG9148 - *supercoiling factor* (*scf*)
CG11798 - *charlatan* (*chn*)		CG6207	CG9171
CG13651 - *distal antenna-related* (*danr*)		CG9005	CG9415 - *X box binding protein-1* (*Xbp1*)
CG13894		CG9238	CG9520
CG18375		CG10130 - *Sec61b*	CG10072 - *sugarless* (*sgl*)
CG31607		CG10521 - *Netrin-B* (*NetB*)	CG10960
CG32434 - *schizo* (*siz*)		CG10746	CG11415 - *Tetraspanin 2A* (*Tsp2A*)
		CG11051	CG11546 - *kermit*
		CG11100	CG15211
		CG11188	CG17932 - *Ugt36Bc*
		CG12177	CG32858 - *singed* (*sn*)
		CG13037 - *mitochondrial ribosomal protein S34* (*mRpS34*)	
		CG15162 - *Misexpression suppressor of ras 3* (*MESR3*)	
		CG31150	
		CG32372	
		CG32423 – *alan shepard* (*shep*)	
		CG33099	
14	7	24	18

Table 5.3. Summary of the results of the reciprocal BLAST search for direct orthologues of the *Drosophila* genes with expression patterns relating to the intercalary segment. Genes are grouped according to whether the BLAST search of the *Tribolium* genome recovered no similar sequences or multiple similar sequences, or whether the reciprocal BLAST search of the *Drosophila* protein database was unable to distinguish a direct orthologue from potential paralogues or did identify a direct orthologue. For the genes where a direct orthologue was identified in the *Tribolium* genome, genes where primers were not designed are shown in italics. The number of genes in each category is shown at the bottom. For more details see section 5.4.2.

No similar sequence	Multiple similar sequences	Potential paralogues	Direct orthologues
CG5059	CG3879	CG1942	CG1322
CG7271	CG6117	CG3097	CG1444
CG9005	CG10960	CG3424	CG3184
CG10746		CG3597	CG3732
CG11051		CG3762	CG4261
CG11100		CG5663	CG4280
CG12708		CG6096	CG4322
CG13894		CG8036	CG4501
CG15211		CG9171	CG5249
CG17786		CG10521	CG5575
CG31607		CG13651	CG5840
CG31629		CG17932	CG5893
CG33099			CG6207
			CG9148
			CG9238
			CG9415
			CG9520
			CG10072
			CG10130
			CG11188
			CG11208
			CG11415
			CG11546
			CG11798
			CG12177
			CG13037
			CG13475
			CG15162
			CG18375
			CG31150
			CG31811
			CG32372
			CG32423
			CG32434
			CG32858
13	3	12	35

No similar sequence

For 13 of the 63 *Drosophila* genes, no *Tribolium* sequence could be identified which showed similarity to the *Drosophila* query sequence. At best *Tribolium* sequences could be found that showed similarity to only a small fraction of the *Drosophila* query sequence, and the quality of alignment was poor (a BIT score of less than 80). The BLAST search returned similar sequences in the *Tribolium* genome for the remaining 50 genes.

Mutliple similar sequences

For three genes, a large number of *Tribolium* sequences showed high levels of similarity to the *Drosophila* query sequence, making it impractical to extract and process all the sequence fragments for the reciprocal BLAST search of the *Drosophila* protein database. Therefore, these genes with multiple similar sequences were set aside from the analysis for practical reasons.

Potential paralogues

It was not possible to distinguish between a direct orthologue and a potential paralogue using the reciprocal BLAST search for a further 12 of the genes. Either the *Tribolium* sequence with the highest E-value recovered genes in *Drosophila* other than or as well as the original query sequence, or multiple highly scoring *Tribolium* sequences recovered the original *Drosophila* query sequence. In these cases it was difficult to ascertain which *Tribolium* sequence was orthologous to the original *Drosophila* sequence, if indeed any direct orthologue existed. Therefore, these genes were also excluded from the analysis.

Direct orthologues

For the remaining 35 genes, the reciprocal BLAST recovered a clear *Tribolium* orthologue. Primers were designed to PCR amplify partial cDNAs of these genes for

probe synthesis. For three of these genes primers were not designed and so partial cDNAs were not amplified.

5.4.3 Tribolium *expression patterns*

Expression patterns were examined for the 32 genes with orthologues in *Tribolium* for which primers were synthesised. No clear *Tribolium* expression pattern could be seen for 15 of these genes. Either no expression pattern was clearly visible, or the embryo stained strongly with background. A variety of expression patterns were seen in the remaining 17 genes, not all relating to the intercalary segment. These are presented in figure 5.3. I will now describe the *Tribolium* expression patterns, illustrating where they appear to be conserved with *Drosophila*. Table 5.4 summarises which genes had localised expression patterns and which of these had expression patterns associated with intercalary segment.

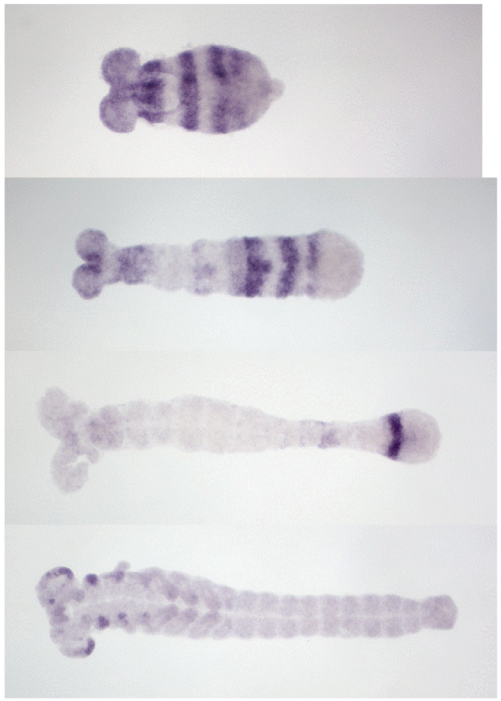
Ectodermal expression

Four of the genes with *Tribolium* expression patterns had *Drosophila* orthologues where expression was associated with the posterior of the procephalon (figure 5.3 A). One of these genes (*Tc*-CG32434) showed no expression associated with the posterior of the procephalon in *Tribolium*. Instead, there appeared to be expression associated with the middle plate and the prospective mesoderm, although this was not very striking. The other three genes all showed a band of expression across the anterior of the embryo. *Tc*-CG5249 and *Tc*-CG18375 both showed a band of expression immediately posterior to the head lobes, where gene expression associated with the intercalary segment would

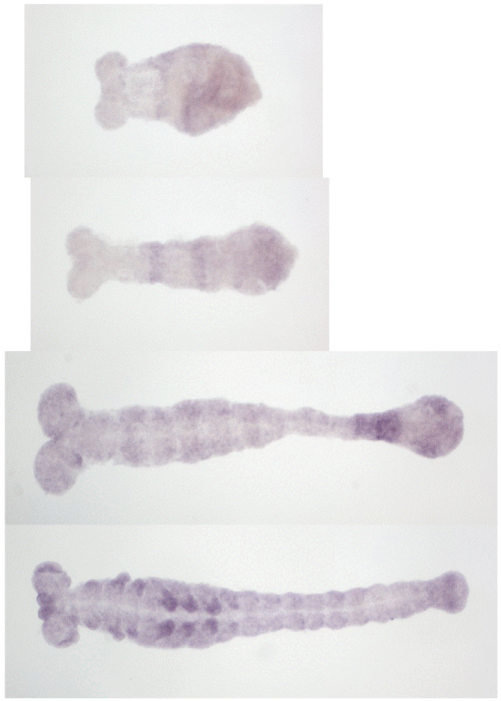
Figure 5.3. *Tribolium* expression patterns for orthologues of the genes with expression patterns relating to the *Drosophila* intercalary segment. *(Following page).*. Embryos orientated with anterior to the left. Brightfield images. For each gene, expression patterns are shown around the time of gastrulation and at early, middle and late times in germband extension. For *Tc*-CG4501 and *Tc*-CG11415, expression in germband retracting embryos are also shown as there is no obvious expression during earlier stages. Genes are presented grouped in accordance with where their *Drosophila* orthologue was expressed, namely the posterior of the procephalon (A), other ectodermal expression domains (B), early mesodermal expression (C) and late mesodermal expression (D). Expression patterns are not shown for genes that did not have any localised expression.

A - "Posterior procephalon" group genes

Tc-CG5249

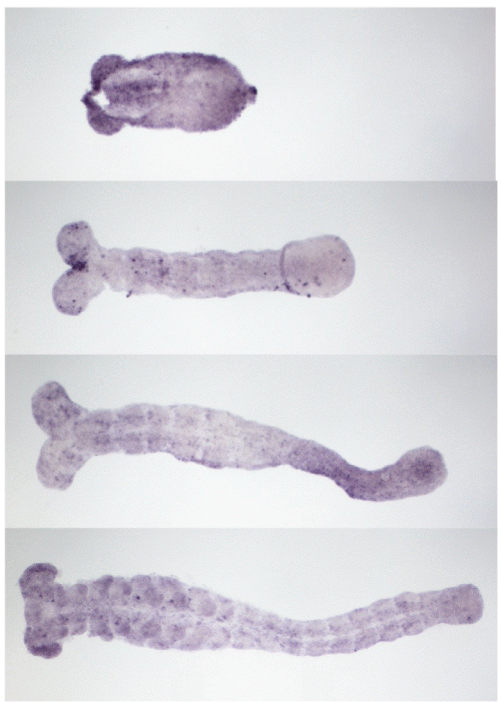


Tc-CG5575



Tc-CG18375

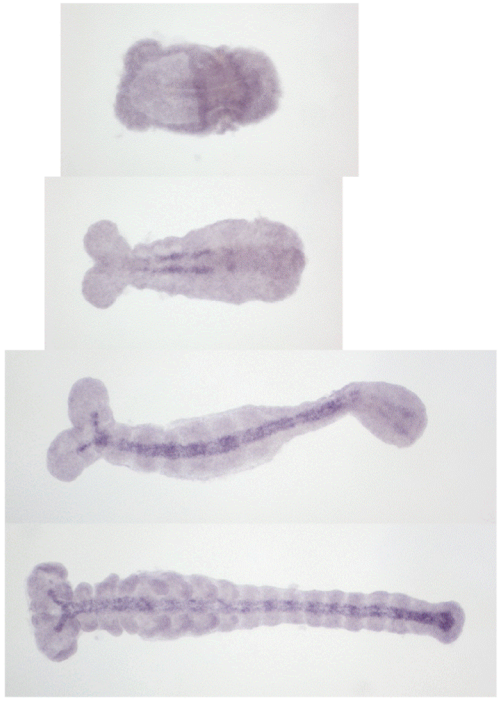
Tc-CG32434



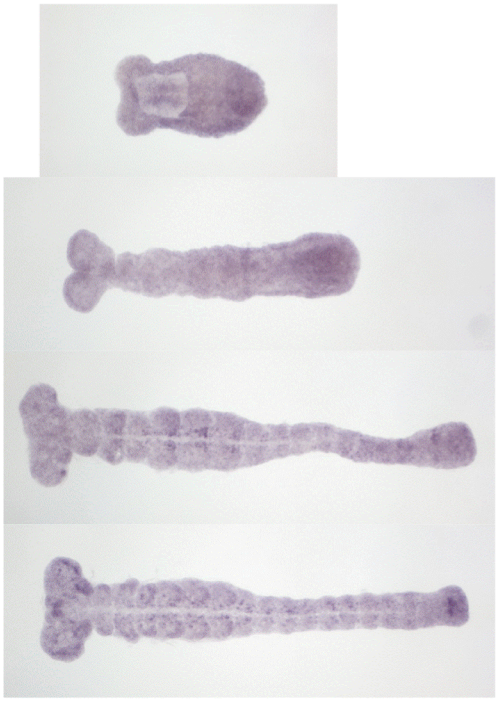
B - "Other ectodermal expression" group genes

Tc-CG5893

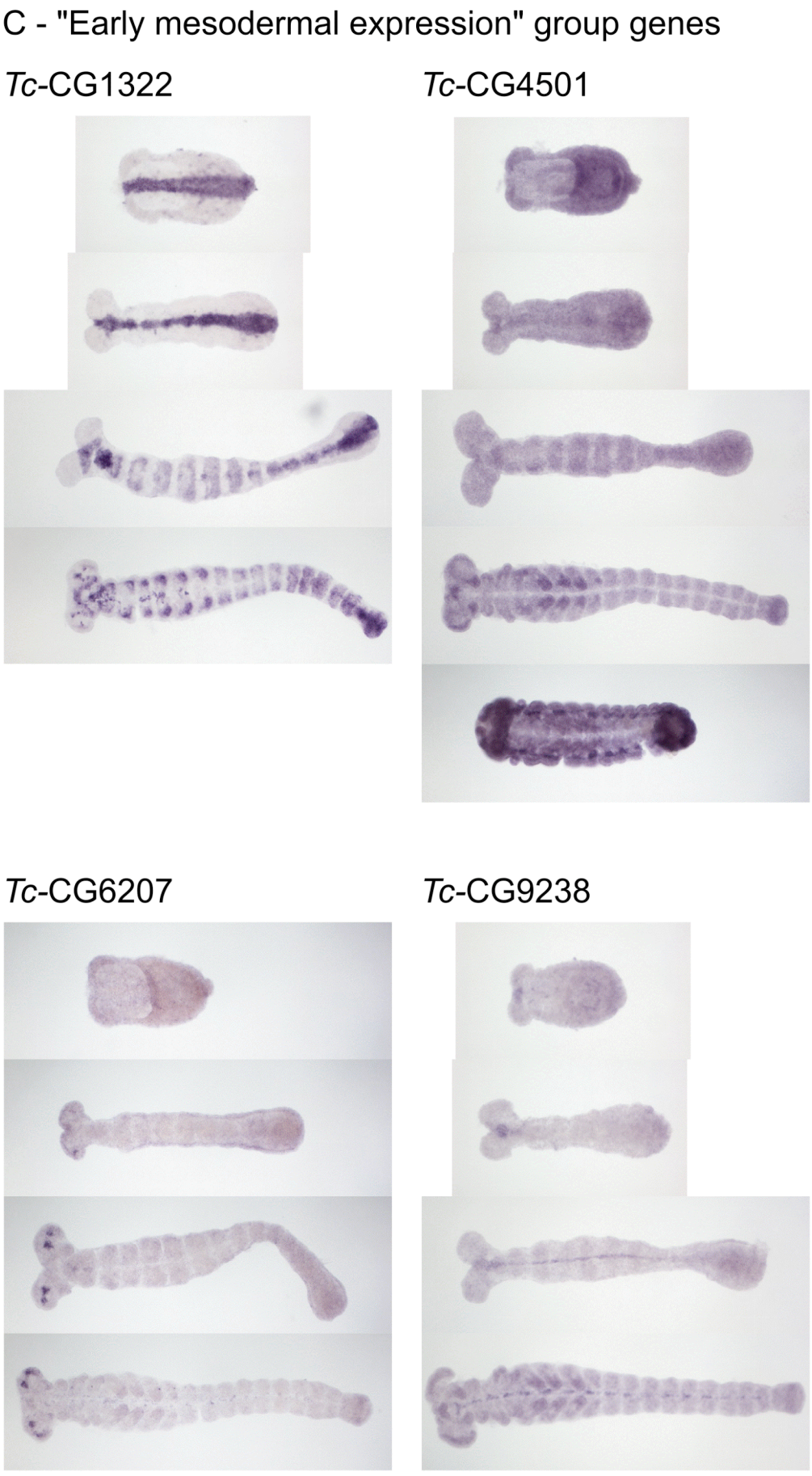
Tc-CG13475



Tc-CG31811



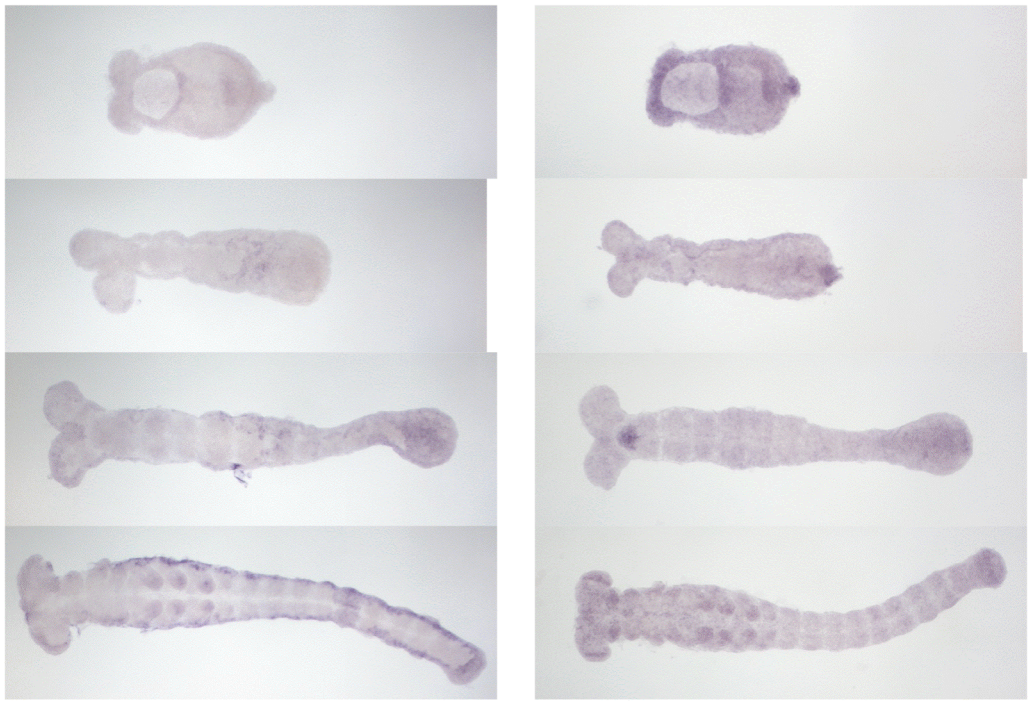
(Figure 5.3 continued)



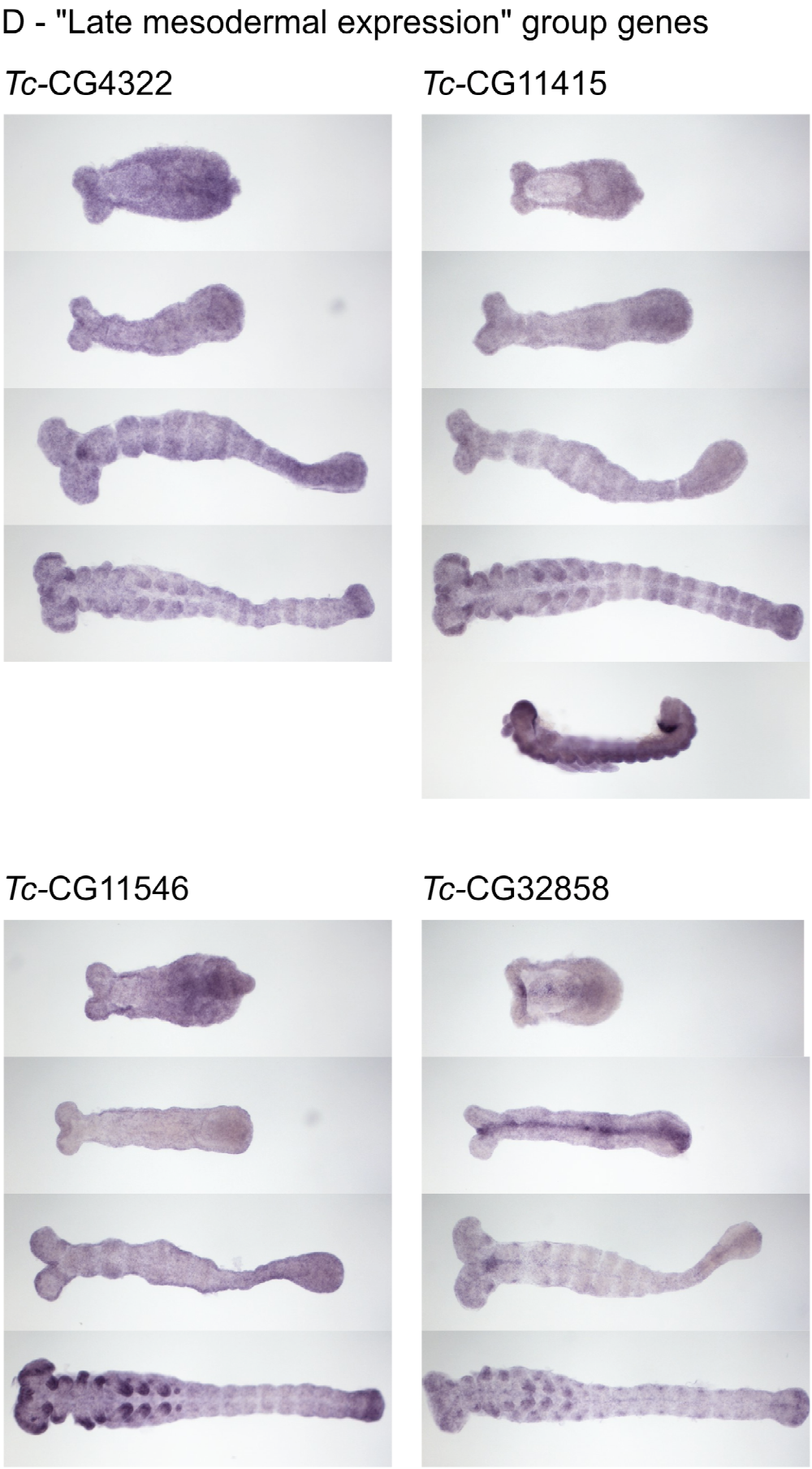
(Figure 5.3 continued)

C (continued)

Tc-CG32372 *Tc*-CG32423



(Figure 5.3 continued)



(Figure 5.3 continued)

Table 5.4. Summary of *Tribolium* expression patterns. Genes are organised according to the expression patterns of their *Drosophila* orthologues (see table 5.2) and are grouped according to whether a detailed examination showed expression associated with the intercalary segment (see section 5.4.4), or there was a localised expression pattern elsewhere in the embryo, or there was no localised expression.

Ectodermal		Mesodermal	
Posterior procephalon	Other	Early	Late
Expression associated with the intercalary segment			
Tc-CG5249		*Tc*-CG32423	*Tc*-CG4322
			Tc-CG32858
Other localised expression patterns			
Tc-CG5575	*Tc*-CG5893	*Tc*-CG1322	*Tc*-CG11415
Tc-CG18375	*Tc*-CG13475	*Tc*-CG4501	*Tc*-CG11546
Tc-CG32434	*Tc*-CG31811	*Tc*-CG6207	
		Tc-CG9238	
		Tc-CG32372	
No localised expression			
Tc-CG3732		*Tc*-CG3184	*Tc*-CG1444
Tc-CG11208		*Tc*-CG4261	*Tc*-CG9148
Tc-CG11798		*Tc*-CG4280	*Tc*-CG9520
		Tc-CG5840	*Tc*-CG10072
		Tc-CG10130	
		Tc-CG12177	
		Tc-CG13037	
		Tc-CG31150	

be expected. For *Tc*-CG5575 the band of expression lay too posteriorly from the head lobes to be implicated in the intercalary segment.

A further three genes with *Tribolium* expression patterns had *Drosophila* orthologues with other expression patterns in the intercalary ectoderm (*Tc*-CG5893, *Tc*-CG13475 and *Tc*-CG31811; figure 5.3 B). None of these three genes appeared to show intercalary segment specific expression in *Tribolium*. *Tc*-CG31811 had localised expression in the head, but in head lobes not the intercalary segment. *Tc*-CG5893 and *Tc*-CG13475 had segmentally repeating expression patterns.

Mesodermal expression

The remaining 10 genes with *Tribolium* expression patterns had *Drosophila* orthologues with expression in the potential intercalary mesoderm (figure 5.3 C and D). 6 of these genes do not appear to show any expression associated with the intercalary segment. Three of them (*Tc*-CG6207, *Tc*-CG9238 and *Tc*-CG11415) were expressed in the head, but this expression did not appear to relate to the intercalary segment. The remaining four genes (*Tc*-CG1322, *Tc*-CG4322, *Tc*-CG32423 and *Tc*-CG32858) all showed a striking domain of expression coincident with the posterior-most extent of the head lobes; the region associated with the intercalary segment. Whilst *Tc*-CG1322 and *Tc*-CG32858 also had expression in other regions of the embryo, expression of *Tc*-CG4322 and *Tc*-CG32423 was restricted to this domain in the head.

5.4.4 Detailed examination of the candidate intercalary segment genes

I further examined the *Tribolium* expression patterns for the genes that appeared to show conserved expression associated with the intercalary segment to confirm whether or not there was in fact expression in the segment.

Ectodermal expression

Tc-CG18375 is expressed posterior to the head lobes. However, closer inspection of its expression showed that whilst there is a band of expression at the back of the procephalon, this is in fact associated with the mandibular segment (figure 5.4 A and B). Additionally expression is restricted to the mesoderm (figure 5.4 C), not the ectoderm as was originally expected.

Tc-CG5249 (see figure 5.4 D-G) has a complicated expression pattern. Expression is seen in the germ rudiment in a series of bands, including one immediately posterior to the head lobes, as well as further expression in the prospective head mesoderm and across the head lobes (figure 5.4 D). These bands persist through gastrulation (figure 5.4 E). As the germband extends, transcripts are maintained immediately posterior to

Figure 5.4. Gene expression at the posterior of the *Tribolium* procephalon. (A-C) Expression of *Tc*-CG18375. (D-G) Expression of *Tc*-CG5249. (A, B, D-G) Ventral views of embryos orientated with anterior up. (C) Lateral view of embryos orientated with anterior left. Nomarski images. *Tc*-CG18375 expression is first seen during early germband extension as a faint band across the embryo posterior to the head lobes (A). As the germband extends further and the segments become morphologically distinguishable, expression intensifies and appears to reside in the mandibular segment (B; arrowheads mark the approximate position of the intercalary-mandibular segment boundary). Additionally, expression appears restricted to the mesoderm (C). *Tc*-CG5249 expression is first seen in the germ rudiment as a series of bands across the embryo; one immediately posterior to the head lobes, with a further two bands in the trunk and faint expression across the head lobes (D). This pattern persists through gastrulation (E). As the germband extends (F), the band of expression at the posterior of the head lobes perisits, whilst the expression across the head lobes begins to fade and the expression in the trunk undergoes a complicated series of modulations. As the segments become morphologically distinct (G), the expression at the posterior of the head lobes can be seen to largely reside in the mandibular segment, extending into the posterior intercalary segment (arrowheads in G mark the approximate position of the intercalary-mandibular segment boundary).

the head lobes, whilst the more posterior bands undergo a series of complex modulations (figure 5.4 F). The band of expression at the base of the head lobes is still apparent when the segments become morphologically distinct (figure 5.4 G). At this point, it largely appears to be associated with the mandibular segment. However, the anterior-most limit of expression appears to extend into the intercalary segment. Therefore, expression of *Tc*-CG5249 does appear to be associated with the intercalary segment.

Mesoderm

Tc-CG1322, *Tc*-CG4322, *Tc*-CG32423 and *Tc*-CG32858 all showed a striking central domain of expression coincident with the posterior-most extent of the head lobes; a region that appears to correspond to at least part of the intercalary segment. However, closer inspection of *Tc*-CG1322 suggests that the expression of this gene is not localised to the intercalary segment but rather is expressed throughout the mesoderm (figure 5.5). The stronger expression associated with the intercalary segment appears to be because this block of mesoderm does not show the typical spreading seen in the mesodermal somites of the other segments, and so whilst expression in mesodermal tissue has thinned out in other segments, it remains as a large block beneath the intercalary segment (figure 5.5 D and E). This block of mesoderm does eventually spread (figure 5.5 F), but even at this stage does not appear typical for a mesodermal somite.

Expression for the other three genes (*Tc*-CG4322, *Tc*-CG32423 and *Tc*-CG32858) is localised to the domain coincident with the posterior of the head lobes (figure 5.6), which would appear to correspond to the intercalary mesoderm (compare with the intercalary segment mesoderm in figure 5.5). *Tc*-CG32858 does have a more extensive early expression pattern (as shown in figure 5.3), but by late germband extension this has become restricted to the domain coincident with the posterior of the head lobes (figure 5.6 G-H). Expression in this domain does not persist past the end of germband extension. *Tc*-CG4322 and *Tc*-CG32423 do not show extensive early expression patterns (as shown in figure 5.3). For *Tc*-CG32423 (figure 5.6 C-F), transcripts begin to accumulate early during germband extension, and persist until late germband extension. In contrast, expression of *Tc*-GC4322 (figure 5.6 A and B) is very transient late in

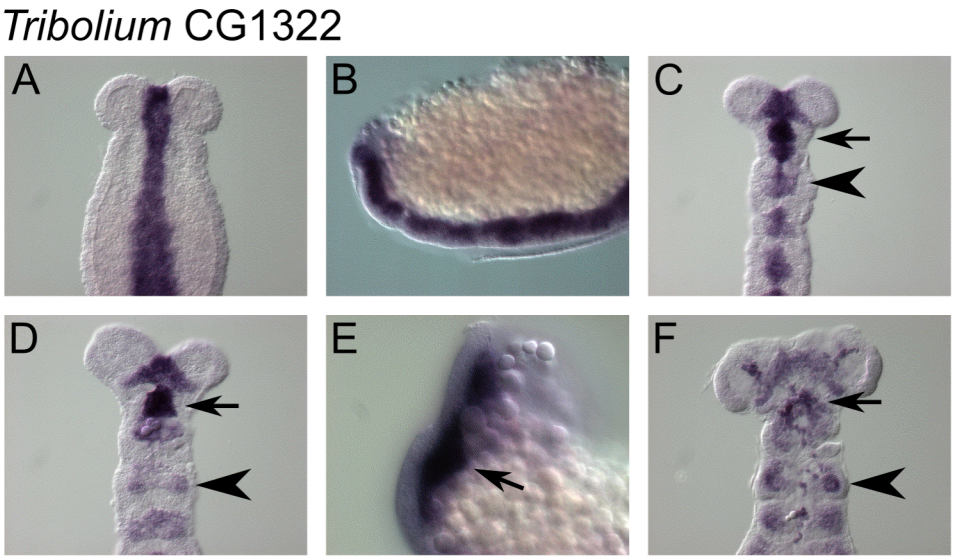


Figure 5.5. Expression of *Tribolium* CG1322. (A, C, D, F) Ventral views of embryos orientated with anterior up. (B, E) Lateral view of embryos orientated with anterior left. Nomarski images. *Tc*-CG1322 is expressed through out the developing mesoderm. Transcripts are first found along the middle plate in the germ rudiment (A), and then are restricted to a deeper layer of the embryo after gastrulation (B). As mesodermal somites form during early germband extension, the expression pattern then breaks into repeated units (C) and spread laterally as the germband further extends (D), before associating with the forming appendages (F). This is exemplified by the maxillary segment (marked with an arrowhead in C, D and F). Expression associated with the intercalary segment (the domain between the antennal and mandibular mesoderm, marked with an arrow in C, D, E and F) remains intense during germband extension (C and D) as cells expressing the gene remain in a large domain and do not spread out laterally (block of cells indicated by arrow in E). By late germband extension (F) this intercalary domain begins to break down and spread laterally.

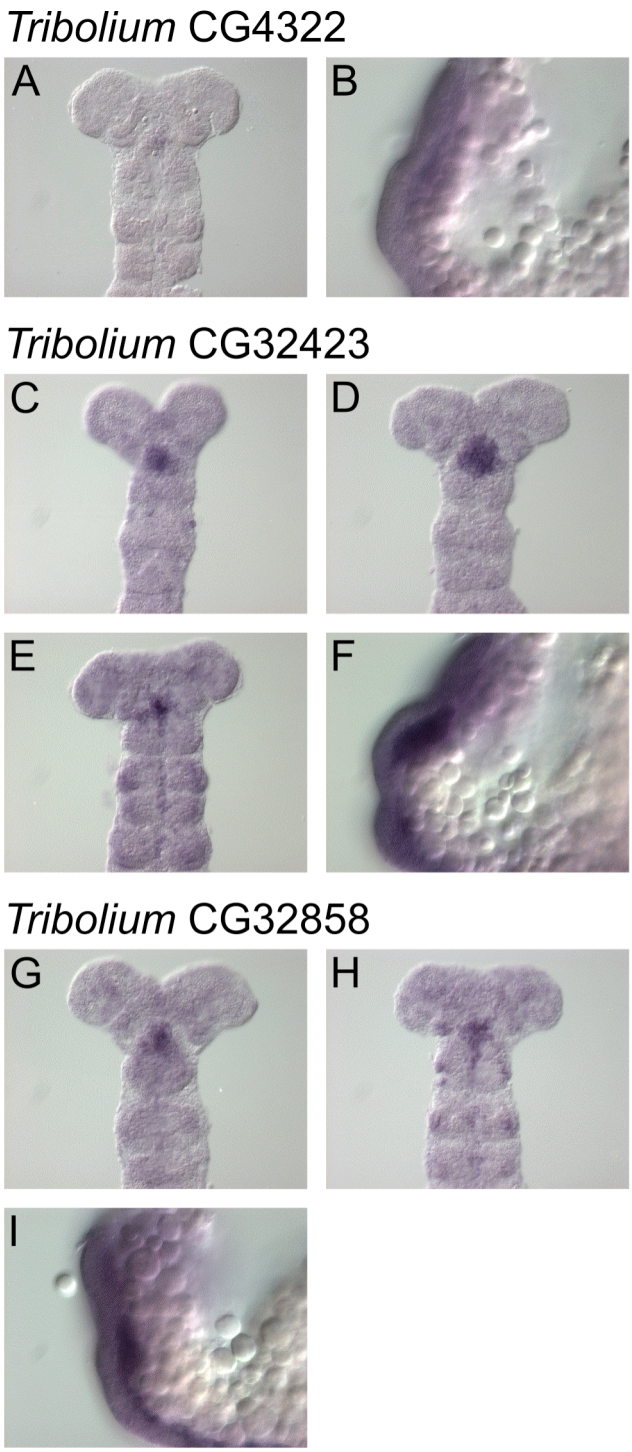


Figure 5.6. Gene expression in the *Tribolium* intercalary segment mesoderm. (A, B) Expression of *Tc*-CG4322. (C-F) Expression of *Tc*-CG32423. (G-I) Expression of *Tc*-CG32858. (A, C-E, G, H) Ventral views of embryos orientated with anterior up. (B, F, I) Lateral view of embryos orientated with anterior left. Nomarski images. *Tc*-CG4322 is expressed in a central domain coincident with the posterior-most extent of the head lobes transiently late in germband extension (A). Expression is restricted to the mesoderm (B). *Tc*-CG32423 is expressed in a central domain coincident with the posterior-most extent of the head lobes from early in germband extension (C). As germband extension continues this domain of expression intensifies (D) before reducing in size by late germband extension (E). Expression is restricted to the mesoderm (F). *Tc*-G32858 expression is restricted to a central domain coincident with the posterior-most extent of the head lobes in the later stages of germband extension (G, H). Expression is restricted to the mesoderm (I).

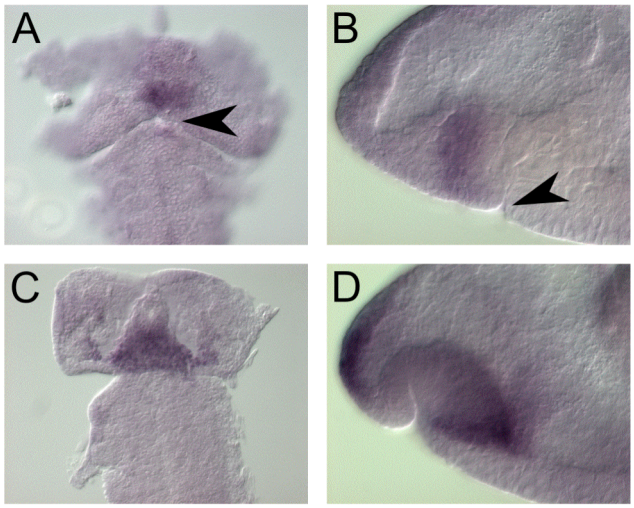
germband extension. Importantly, for all three genes, expression in this domain appears restricted to the mesoderm, being localised in a layer of the embryo beneath the ectoderm.

5.4.5 Expression of mesodermal genes in Drosophila

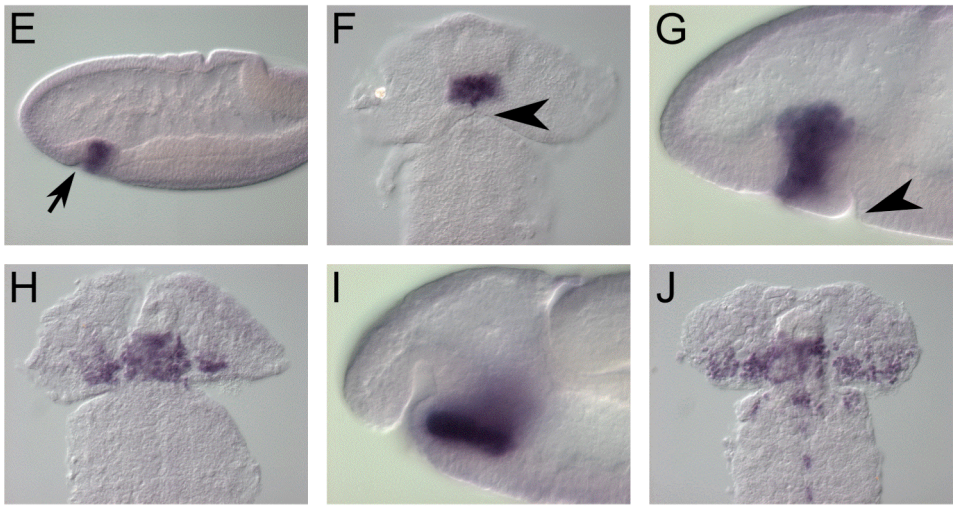
The three genes identified as expressed in the *Tribolium* intercalary mesoderm were expressed in a single central domain. In other segments mesoderm spreads laterally to form somites. Therefore the *Tribolium* expression patterns appear atypical for the mesoderm. I examined the expression patterns of the three genes in detail in *Drosophila* (figure 5.7), where the development of the head mesoderm has been described (de Velasco*, et al.*, 2006).

CG4322, CG32423 and CG32858 are all expressed in what de Velasco *et al.* (2006) describe as the intercalary mesoderm. All three genes are first seen at the anterior of the ventral furrow – anterior to the cephalic furrow or behind the "T-bar" (figure 5.7 A, B, E-G, K and L) – with expression subsequently spreading laterally along the posterior of the procephalon (figure 5.7 C, H, J, M and O). There are differences in the relative timings of expression. CG32423 is expressed first, with transcripts accumulating in the gastrulating (stage 7) embryo (figure 5.7 F). Expression of CG4322 and CG32858 is first seen in the germband extending (stage 8) embryo (figure 5.7 A and K). However, expression of CG4322 is most transient with transcripts not seen after late germband extension (stage 9) (figure 5.7 C and D), whilst in the germband extended (stage 11) embryo CG32423 and CG32858 are expressed extensively across the back of the procephalon (figure 5.7 J and O).

Drosophila CG4322



Drosophila CG32423



Drosophila CG32858

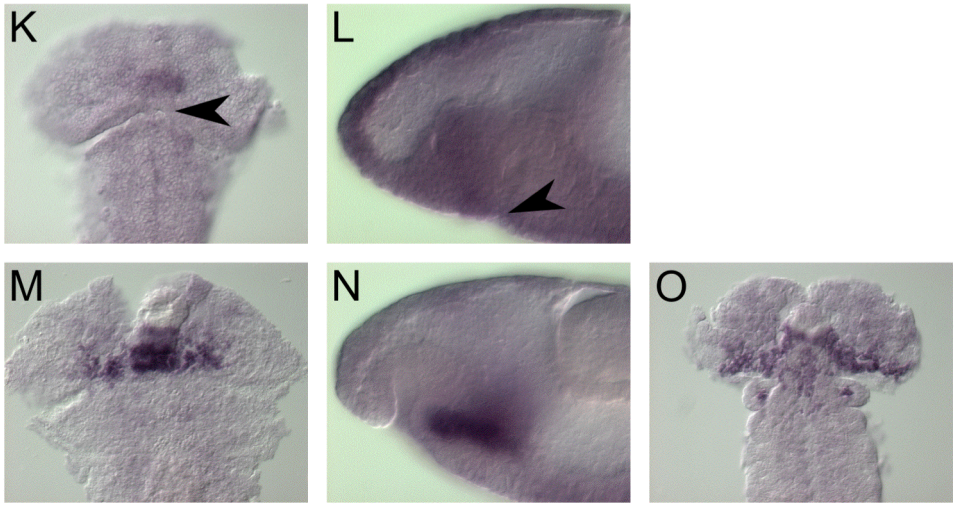


Figure 5.7. Gene expression in the *Drosophila* intercalary segment mesoderm. *(Previous page).* (A,-D) Expression of CG4322. (E-J) Expression of CG32423. (K-O) Expression of CG32858. (A, C, F, H, J, K, M, O) Ventral views of embryos orientated with anterior up. (B, D, E, G, I, L, N) Lateral view of embryos orientated with anterior left. Nomarski images. CG4322 expression in first seen in the germband extending (stage 8) embryo in a central domain immediately anterior to the cephalic furrow (cephalic furrow marked by arrowhead in A and B). Later in germband extension (stage 9) the domain of expression has begun to spread laterally (C). By this stage expression is clearly restricted to the mesoderm (D). CG32423 expression is first seen in the gastrulating (stage 6) embryo at the anterior of the ventral furrow (arrow in E marking the position of the T-bar). In the germband extending (stage 8) embryo (F, G) this domain is seen to lie immediately anterior to the cephalic furrow (cephalic furrow marked by arrowhead in F and G). Later in germband extension (stage 9) this domain of expression has begun to spread laterally (H) and is restricted to the mesoderm (I). In the germband extended (stage 11) embryo, expression has spread across the posterior of the procephalon (J). CG32858 expression is first seen in the germband extending (stage 8) embryo (K, L) in a central domain immediately anterior to the cephalic furrow (cephalic furrow marked by arrowhead in K and L). Later in germband extension (stage 9) this domain of expression has begun to spread laterally (M) and is restricted to the mesoderm (N). In the germband extended (stage 11) embryo, expression has spread across the posterior of the procephalon (O).

5.5 Discussion

The screen described in this chapter was designed to identify genes with conserved expression in the intercalary segment between *Drosophila* and *Tribolium*. The aim of the screen was to identify potential candidate genes for patterning the intercalary segment. Four genes were recovered: CG5249 in the posterior intercalary segment ectoderm and CG4322, CG32423 and CG32858 in the intercalary segment mesoderm.

5.5.1 Methodological factors contributing to a lack of conservation

This screen was based on the assumption that there would be a number of genes involved with patterning the intercalary segment across the insects, and that any such genes would have conserved expression associated with the segment. It is, therefore, surprising that only four genes with potential intercalary segment expression patterns were recovered from an original set of 63 *Drosophila* candidates identified in the BDGP expression pattern database. The small proportion of genes with a conserved expression pattern suggests that the developmental processes involved in patterning the segment

are not conserved between the two insects. However, a number of methodological factors could have contributed to this discrepancy.

Selection of Drosophila *candidate genes*

The first step in the screen protocol was to select *Drosophila* gene expression patterns associated with the intercalary segment, from the BDGP expression pattern database. It is possible that several of these candidate genes were not in fact expressed in the *Drosophila* intercalary segment. In some cases, the proposed intercalary segment expression may have been an artefact. For example, two of the areas of interest – the back of the procephalon and the intercalary mesoderm – appear to be associated with thicker layers of embryonic tissue resulting from the various furrows that form during gastrulation. The thicker tissue could make background or a more widespread expression pattern look like more intense staining in these areas. This could explain the apparent intercalary segment expression for genes like CG1322, CG3184 and CG3732. In fact, CG1322 has been studied in *Drosophila* where it is known as *zinc-finger homeodomain protein 1* (*zfh-1*) (Lai*, et al*., 1991) and it is expressed throughout the mesoderm. If the proposed intercalary expression patterns of these genes are interpreted as artefacts, then the expression patterns are in fact conserved in *Tribolium*. *Tc*-CG3184 and *Tc*-CG3732 showed no localised expression in the beetle, whilst *Tc*-CG1322 is expressed throughout the mesoderm.

Other genes had striking *Drosophila* expression patterns, with features in the head which were interpreted as belonging to the intercalary segment, and several aspects of these expression patterns appeared to be conserved in *Tribolium*. However, whilst the expression in the head did appear to be conserved, it was clearly not localised to the intercalary segment. This suggests that the original *Drosophila* expression pattern was misidentified as being associated with the intercalary segment. For example, in both insects CG13475 is expressed along the embryonic midline and whilst *Drosophila* has two distinctive domains in the head, in *Tribolium* this is clearly just a bifurcation of the trunk domain of expression. In the light of these probable expression artefacts and misidentifications, it seems that the levels of conserved gene expression patterns

between *Drosophila* and *Tribolium* are higher than four genes out of 63, although not all the conserved genes relate to the intercalary segment.

Genes were misidentified as being expressed in the intercalary segment as a result of the way they were chosen from the BDGP expression pattern database. Selection of *Drosophila* genes was deliberately naïve. Genes with questionable intercalary segment expression patterns were included to ensure that no true intercalary genes were missed. The *in situ* hybridisations for the candidate genes could have first been repeated in *Drosophila* before looking at *Tribolium*, perhaps with segmental marker genes. This would have confirmed whether or not a gene was expressed in the *Drosophila* intercalary segment. However, it made more sense to investigate the beetle first, as this addressed the vital issue of conservation as well as whether there was expression in the intercalary segment. For the same reason an extensive investigation of the *Drosophila* literature was not undertaken at the start of the screen, even though such a review would have shown that genes such as CG1322 (*zfh-1*) do not have localised expression in the *Drosophila* intercalary segment (Lai, *et al*., 1991).

Removal of genes from the dataset

In situ hybridiation was only carried out for *Tribolium* genes where direct orthology could be established with *Drosophila*; by definition an expression pattern can only be conserved in an orthologue. For a number of genes direct orthologues could not be identified either because too many similar sequences were identified in the *Tribolium* genome for the reciprocal BLAST search to be practical, or because the reciprocal BLAST procedure could not distinguish a direct orthologue from possible paralogues. However, problems preventing the identification of an orthologue do not mean that no orthologue was present in *Tribolium*. It is possible that the genes discarded from the screen do have a conserved intercalary segment expression pattern.

Similarly, primers were not designed for three of the genes where direct orthologues had been identified. However, primers could have been designed and probes synthesised. These genes could have conserved expression patterns in the intercalary segments of *Drosophila* and *Tribolium* and this could be examined. Once the various

stages where genes were discarded for practical reasons are accounted for, the pool of *Drosophila* genes is reduced from 63 to 45 (three genes had multiple similar sequences, for 12 genes direct orthologues could not be distinguished from possible paralogues and for three genes primers were not designed).

In situ *hybridisation conditions*

A number of genes with clear *Drosophila* expression pattern showed no localised expression in *Tribolium*. It is possible that this is a true reflection of gene expression in *Tribolium*. However, all probes were synthesised from cDNA so there must have been some level of expression at some point during embryogenesis. It is possible that for several of the genes in *Tribolium* there was localised expression, but this was not detected due to problems with the probes. For example, some probes may not have been sensitive enough to detect a potentially weak expression pattern.

There are a number of ways *in situ* hybridisation conditions could be optimised to try to address such practical problems, for example a range of probe concentrations could have been tried for each gene. Also, different probes could have been synthesised for each gene by amplifying different partial cDNAs. However, these measures were not practical when screening through a large set of genes. Therefore, it is possible that some genes with conserved expression were missed because the expression pattern was not visualised.

5.5.2 The level of conserved expression between Drosophila *and* Tribolium

Various methodological factors could, therefore, have contributed to the discrepancy between the original number of candidate genes in *Drosophila*, and the number with a conserved expression pattern in *Tribolium*. However, there are clearly genes where the *Drosophila* expression pattern was not conserved in *Tribolium*. A number of genes showed obvious expression around the *Drosophila* intercalary segment, but had different expression patterns in *Tribolium*. For example CG4501 shows striking expression in the head mesoderm of *Drosophila* as well as segmentally repeated

expression along the trunk of the germband retracting embryo. In the *Tribolium* orthologue only the trunk expression was observed.

Also, whilst it is true that for the genes where a direct orthologue could not be indentified there may still have been a *Tribolium* orthologue, there is no reason to expect that this would be the situation. In many cases these paralogy relationships could have involved gene duplications and deletions between the two insects and so it is possible that the *Drosophila* intercalary candidate genes were novel genes, or the *Tribolium* orthologues were lost. It is noteworthy that in the BLAST search of the *Tribolium* genome, a small group of genes showed no sequence similarity in *Tribolium*. This may have been because the *Tribolium* genome sequence was not complete. However, it is more likely that these were genuine examples where there was no *Tribolium* orthologue. The split between the lineages leading to *Drosophila* and *Tribolium* is ancient – fossil beetle remains are known from the Permian (over 250 million years ago) (Lubkin and Engel, 2005). Moreover, comparisons of the *Tribolium* genome with other insect and vertebrate genomes show that whilst 15% of the predicted 16,404 *Tribolium* genes have universal single copy orthologues and 9% have insect specific orthologues, "thousands" of genes appear to be species specific, with no orthologue in *Drosophila* or any genome examined (Richards*, et al.*, 2008).

Unfortunately, the various methodological factors make it difficult to give a proportion of genes which have conserved expression. However, it is clear that whilst some features of development are conserved, others are not. This emphasises the importance of comparing across organisms when trying to understand how a bodyplan character is patterned. It is of interest to know what (if any) aspects of intercalary segment patterning the genes with conserved expression are involved with, and indeed whether there is any reason why some aspects of development are more likely to vary between closely related organisms than others. Further work looking at the different functions of the genes expressed in the intercalary segment should begin to shed light on this.

5.5.3 Implications for the development of the intercalary segment

The screen protocol identified four genes with conserved expression patterns associated with the intercalary segment. In the absence of functional work nothing can be said about any precise role the genes have in the development of the intercalary segment or its evolution. However, there are some important points which can be made based on the expression patterns.

Drosophila *and the function of CG5249*

CG5249, for which the *Tribolium* orthologue is expressed in the ectoderm overlapping the posterior of the intercalary segment, has previously been studied in *Drosophila* where it is known as *Blimp-1* (Ng, *et al.*, 2006). The early *Drosophila* expression pattern appears conserved with *Tribolium*. There are three bands of expression with one at the posterior of the procephalon and there is further expression in the anterior head. Moreover, the modulations seen in the beetle and the fly are remarkably similar, with the expression at the posterior of the procephalon persisting after the rest of the expression has mostly faded. Further *in situ* hybridisations with markers are required to see the extent to which this expression at the back of the procephalon is conserved.

Ng *et al*. (2006) comment that the expression of *Blimp-1* is reminiscent of a gap gene. Indeed, the expression at the back of the procephalon is similar to the various genes involved with the development of the intercalary segment such as *kn* or the head "gap-like" genes of *Drosophila*. Like these genes it also encodes a transcription factor (containing a zinc finger and SET/PR domain). However, in *Drosophila Blimp-1* function has been studied and RNAi shows the gene has a role in patterning the tracheae. There was no obvious early gap-like phenotype. The authors suggest that the gene is dispensable for segmentation. If this is the case, then the gene cannot have a conserved function in patterning the intercalary segment between the beetle and fly. However, as illustrated in chapter 1, several studies have shown the early head development in *Drosophila* has many derived features. The *Drosophila* head gap genes are not conserved in *Tribolium* in terms of either expression or function. Therefore, it is still interesting to investigate the function of the *Tribolium* orthologue of *Blimp-1* (*Tc-*

CG5249), to see if there is any function affecting the intercalary segment in the beetle. If there were, it would be a suitable candidate gene for further study in other insects.

Gene expression in the mesoderm of Drosophila *and* Tribolium

Drosophila and *Tribolium* ortholgues of CG4322, CG32423 and CG32858 all appear to show mesodermal expression associated with the intercalary segment. All three genes have been studied at some level in *Drosophila*: CG4322 is known as *moody* (Daneman and Barres, 2005), CG32423 is known as *alan shepard* (*shep*) (Bjorum, 2006) and CG32858 is known as *singed* (*sn*) (Cant*, et al.*, 1994). However, none of the genes have been implicated in mesodermal development, and the behaviour of the tissue expressing these genes in *Tribolium* appears atypical for mesoderm. In all other segments, mesoderm spreads laterally to form paired coelomic sacs (Handel*, et al.*, 2005). Paired somites have been described in the intercalary segment of other insects, such as the beetle *Tenebrio molitor*, although these are of a derived cell type and form late (Ullmann, 1964). It is, therefore, unclear what part of the mesoderm these genes are expressed in. The anterior midgut of insects also forms from the anterior middle plate although its exact position is unknown in *Tribolium*. Classical fate mapping studies in a range of insects locate the midgut anlage just posterior to the stomodem (Anderson, 1973) and in *Drosophila*, the primordium of the anterior midgut and the intercalary segment mesoderm both form from the B-C regions of the prospective head mesoderm of de Velasco *et al*. (2006). Therefore, it is even conceivable that these genes are not expressed in the intercalary segment mesoderm of *Tribolium* but rather the anterior midgut.

There are many similarities between *Tribolium* and *Drosophila* in timings of expression for these three genes. In both insects *shep* (CG32423) orthologues are expressed earliest and *moody* (CG4322) orthlogues have the most transient expression. The conserved timings suggest conserved expression. However, in *Drosophila* these three genes are clearly expressed in the mesoderm. The expression of transcripts spreads laterally across the posterior of the procephalon, and the expression along the posterior of the procephalon is in the region described as intercalary mesoderm by de Velasco *et al*. (2006). Interestingly, the mesodermal cells deriving from the intercalary segment of

Drosophila are described as crystal cells – a subset of hemocyte cells. Hemocyte development has been studied in the moth *Manduca sexta* (Nardi, 2004). Here stains for granular cells (a type of hemocyte) show them in a central mesodermal domain in the head, behind what appears to be the intercalary segment.

Tc-moody, *Tc-shep* and *Tc-sn* are expressed in a central domain in the *Tribolium* head, resembling the site of formation of hemocytes in *Manduca*. Furthermore, in *Drosophila* the orthologues of these three genes are expressed in what appear to be prospective hemocytes, and their timings of expression are conserved in *Tribolium*. This suggests that the three genes may have conserved expression in the prospective hemocytes of *Tribolium* as well. This is of considerable interest, as de Velasco *et al*. (2006) argue that hemocytes are the major mesodermal derivative of the insect intercalary segment.

It is worth noting that although these three genes have previously been studied in *Drosophila*, they have not been studied to a great extent and so it is probable that they have as yet unknown functions. However, based on what is known about their functions, there are some aspects that could be of potential interest to intercalary segment development. *sn* has been implicated in actin bundle formation, being required for bristle formation and nurse cell cytoplasm transport (Cant, *et al*., 1994). It is not immediately obvious what role such a gene could have in hemocyte formation. *shep* produces a putative RNA binding protein suggesting a possible regulatory role in development, but as yet has only been implicated in gravitaxis (Bjorum, 2006). *moody* produces a G protein-coupled receptor suggesting a possible role in signalling. Moreover, this gene has been implicated in the formation of the blood-brain barrier (Daneman and Barres, 2005) which could be of potential interest to the intercalary segment as hemocytes give rise to blood cells.

5.6 Conclusions

I presented a screen to find new candidate genes for patterning the intercalary segment. Searching for genes with conserved expression patterns in the intercalary segments of the fruit fly *Drosophila melanogaster* and the red flour beetle *Tribolium castaneum* recovered four such genes: one expressed in the posterior intercalary segment ectoderm in a domain reminiscent of the head gap-like genes and *kn*, the other three expressed in the intercalary segment mesoderm in what may be precursors of hemocytes. Given the range of embryonic structures for which expression patterns are annotated in the Berkley *Drosophila* Genome Project expression pattern database, this approach of searching for conserved expression between *Drosophila* and *Tribolium* seems to be a productive method for finding new candidate genes for patterning a range of structures. Also, it is apparent that whilst some genes with expression in the *Drosophila* intercalary segment have conserved expression in the *Tribolium* intercalary segment, others do not. This emphasises the importance of taking a comparative approach when studying the development of a conserved morphological feature; any one organism is likely to have several derived features.

Chapter 6: Discussion

6.1 Overview

In this thesis I set out to investigate the evolution of the insect bodyplan and in particular, the key transition from the crustacean second antennal segment to the intercalary segment of the insect head. I first set out to establish a phylogenetic framework in which to view this transition. I investigated the phylogeny of the Pancrustacea and the position of the insects, finding further evidence for a close relationship between the hexapods and the branchiopod crustaceans (chapter 3). I then concentrated on the development of the intercalary segment. First I addressed the problem of what constitutes the intercalary segment in the embryo of the model system *Drosophila melanogaster*. I presented a detailed comparison of gene expression between the fruit fly and the red flour beetle *Tribolium castaneum*, confirming that the hypopharyngeal lobes of the *Drosophila* embryo do not belong to the intercalary segment as had previously been thought (chapter 4). Then, I presented a screen to find more candidate genes for patterning the intercalary segment, recovering four genes with conserved expression patterns associated with the intercalary segment of *Drosophila* and *Tribolium*: one gene with expression in the posterior intercalary segment ectoderm and three genes with conserved expression in the intercalary segment mesoderm (chapter 5).

6.2 Implications of phylogeny

6.2.1 Inferring ancestral developmental pathways

In chapter 1 I outlined the importance of having an established phylogenetic framework for evolutionary developmental studies, such as the evolution of the intercalary segment. An established phylogeny allows the development of the ancestral and derived character states to be inferred at both ends of the stem lineage. Only in this framework can the developmental changes underlying the morphological transition within this lineage be inferred. In this light, one of the most important results to come from the phylogenetic work presented in chapter 3 is that the branchiopod crustaceans were placed closer to the insects than the malacostracan crustaceans.

The importance of this result becomes apparent when the distribution of crustacean developmental systems is considered. The majority of crustaceans that have proved most amenable to developmental studies are either branchiopods (for example *Artemia franciscana* and *Daphnia pulex* (Copf*, et al.*, 2003, Papillon and Telford, 2007)) or malacostracans (for example *Parhyale hawaiensis*, *Orchestia cavimana* and *Porcellio scaber* (Abzhanov and Kaufman, 1999, Pavlopoulos and Averof, 2005, Wolff and Scholtz, 2006)). In the absence of a phylogenetic framework there would always be a degree of ambiguity in any developmental comparison between an insect and crustacean. For example a developmental comparison between the insects and the branchiopod crustaceans may highlight developmental changes that appear to be associated with a morphological transition. However, branchiopod development could well have several derived features, and therefore may not represent the ancestral state at the base of the insect stem lineage. Making comparisons with the malacostracan crustaceans would remove the ambiguity. As the malacostracans form an outgroup to the insect-branchiopod grouping, if the developmental state is shared between the branchiopods and malacostracans it is likely to represent the ancestral state at the base of the insect stem lineage.

6.2.2 The diversification of the arthropods

The emerging picture of pancrustacean phylogeny should give further insight into how the insect bodyplan evolved. The characters found in the different crustacean groups can be mapped onto the phylogeny allowing the identification of further character transitions involved in the evolution of the insect bodyplan. Knowing the patterns of tagmosis or the appendage types of successive outgroups to the insects should allow the morphological transitions giving rise to the distinctive insect bodyplan to be inferred.

In chapter 1 I illustrated how the different crustacean groups have very different bodyplans, and how in the face of this morphological diversity there was little consensus between different morphology based crustacean phylogenies. In the context of the phylogenetic framework that I have recovered, it is very difficult to find any convincing synapomorphies between these different groups to support any nodes in the tree. This makes it very difficult to establish any morphological transitions involved in the evolution of the insect bodyplan.

It is important to remember that although my analyses gave strong support to a hexapod-branchiopod sister-grouping, the hypothesis tests did show some ambiguity in the signal regarding the position of the branchiopods, in many ways resembling a soft polytomy. This could be the result of a rapid diversification at the base of the Pancrustacea. Interestingly, the hexapods (including the insects) and the major crustacean groups appear to inhabit very different ecological niches. The hexapods are a terrestrial radiation, the branchiopods a freshwater radiation, the copepods have many planktonic forms, the cirripedes are sessile filter feeders and the malacostracans include a diversity of forms living in all aquatic and some terrestrial environments (Brusca and Brusca, 2003). It is, therefore, conceivable that if the Pancrustacea did undergo a very rapid diversification, it was driven by an ecological radiation into these different niches. This could have been coupled with an equally rapid diversification of their morphology to fit these niches, obscuring any morphological synapomorphies.

However, it is also noteworthy that there are several crustacean groups I could not position in my phylogenetic analysis. I could not resolve the positions of the remipedes

and cephalocarids, and there are various poorly known crustacean groups such as the mystaccocarid crustaceans, which were not represented in the analysis, as there is little sequence data. Moreover, there are a number of fossil crustaceans which show little affinity to any of the major groups such as *Cambronatus*, *Wingertshellicus* and *Eschenbachiellus* (Briggs and Bartels, 2001). As all these different groups show a number of bodyplans different to those of the taxa that I was able to position in the tree, it is still possible that a crustacean phylogeny could be established from which the character transitions involved in the evolution of the insect bodyplan could be inferred. Further phylogenetic analyses, both molecular and morphological are needed.

6.3 Patterning the intercalary segment

I now turn to the developmental changes underlying the transition from the crustacean second antennal segment to the insect intercalary segment. I identified several genes with conserved intercalary segment expression patterns between *Drosophila* and *Tribolium* (chapters 4 and 5). Whilst the conserved expression patterns suggest that these are all good candidate genes for further study in *Drosophila* and *Tribolium*, and more widely in the insects, based on expression patterns alone little can be said about their roles in patterning the intercalary segment, and therefore any potential role in the evolutionary transition. For example, as was discussed in chapter 5, CG5249 (*Blimp-1*) may have a conserved expression pattern including the posterior intercalary segment of *Drosophila* and *Tribolium*, but it does not appear to have a role in the development of the *Drosophila* head. However, there is good reason to think that these genes may be of potential importance for understanding the development of several of the derived features of intercalary segment morphology outlined in chapter 1.

6.3.1 **knot** *and the reduction in size of the intercalary segment*

knot (*kn*) shows a conserved expression pattern between *Drosophila* and *Tribolium*. The gene is involved in establishing intercalary segment polarity gene expression in *Drosophila*, suggesting that this may be the case in *Tribolium* too. It is not immediately obvious what role *kn* could have played in the evolution of the intercalary segment from the second antennal segment. Segment polarity gene expression in the intercalary segment is not an insect specific feature; like any segment, the crustacean second antennal segment also has segment polarity gene expression at its posterior boundary (Browne*, et al.*, 2005). As *kn* is involved with the establishment of segment polarity gene expression, there is no reason to expect it to be involved in the evolution of the derived features of the intercalary segment.

However, one of the striking features of the intercalary segment is its reduced size and the general vestigial appearance of the segment. As was pointed out in chapter 1, the size of the intercalary segment polarity gene stripes is also reduced, and the onset of their expression is delayed relative to the other cephalic segments. Potentially, this reduction and retardation of expression could be related to the overall reduction in the size of the segment. Given that *kn* is involved in the regulation of these stripes, it is possible that studying the regulation of segment polarity gene expression through *kn* may give some insight into the overall reduction of the segment.

6.3.2 Hemocytes and the intercalary segment mesoderm

In chapter 5, I identified three genes that appear to have conserved expression in the intercalary segment mesoderm of *Drosophila* and *Tribolium*. Moreover, I proposed that these three genes are expressed in the prospective hemocytes. If they are involved in the differentiation of the intercalary segment mesoderm to this fate, then their role in the evolution of the intercalary segment could potentially be very important. Hemocytes appear to be the major derivative of the intercalary segment mesoderm (de Velasco*, et al.*, 2006). The only other structure that has been argued to derive from the intercalary segment mesoderm is the suboesophagal body, and its intercalary origins are debated

(Roonwal, 1937, Ullmann, 1964). However, this appears to only be a transient embryonic structure and its significance is uncertain.

Also, it is unclear whether the intercalary segment produces any muscle – the major mesodermal derivative of other segments. *twist* (*twi*) is expressed in cells that will differentiate into muscle – high levels of *twi* promote the formation of somatic mesoderm and suppress that of other mesodermal derivatives (Handel, *et al*., 2005) – but in their description of *Tribolium twi*, Handel *et al*. (2005) do not appear to show any expression between the mandibular and antennal mesoderm, namely in the intercalary segment mesoderm.

In contrast, the mesoderm in the crustacean second antennal segment gives rise to typical somites (Anderson, 1973). There does not appear to be any literature on the origin of hemocytes in crustaceans so it is unclear whether these also derived from the second antennal segment. If they are not, then the production of hemocytes from the intercalary segment would be a novelty associated with the evolution of the intercalary segment from the second antennal segment. It is therefore of considerable interest to further study hemocyte development in both insects and crustaceans. The three genes I have identified would seem to be good candidates for investigation.

6.3.3 Intercalary segmental identity

My results, therefore, provide starting points for the investigation of two important features of intercalary segment development: the reduction in size of the segment and the derived fate of its mesoderm. However, the allocation of intercalary segment identity is still unclear. The screen described in chapter 5 recovered no obvious candidate genes. Such a gene would be of interest as it may give insight into the other major morphological feature of the intercalary segment, namely the loss of its appendages. For example the hox gene *abdominal-A* (*abd-A*) has been implicated in allocating segmental identity to the insect abdominal segments, and it appears to repress leg development in both *Drosophila* and *Tribolium* (Lewis, *et al*., 2000, Vachon, *et al*., 1992).

It is not obvious that the existence of any such gene should be expected. The most obvious candidates for such a role are the hox genes, and whilst *labial* (*lab*) is expressed throughout the segment in all insects investigated (including *Drosophila* as I showed in chapter 4), it does not have a role in segmental identity where studied (*Drosophila* and the milkweed bug *Oncopeltus fasciatus*). However, the head gap genes *empty spiracles* (*ems*) and *buttonhead* (*btd*) have been implicated in segmental identity in *Drosophila*. This suggests that there may not be a single gene giving the intercalary segment its identity in the way the hox genes typically do for several segments (Hughes and Kaufman, 2002b). Rather, genes may operate in a combinatorial manner. As the roles that *ems* and *btd* play in *Drosophila* are not conserved in *Tribolium*, it is unlikely that they are conserved more widely in the insects. However, the possibility that the co-expression of gap-like genes may be involved in allocating segmental identity in other insects makes the function of genes with gap-like expression patterns, like *Blimp-1*, of considerable interest.

There is one final issue relating to segmental identity that is worth discussing. *lab* and *proboscipedia* (*pb*) are expressed in the second antennal segment of the crustacean *Porcellio*, and in the homologous segment in other arthropods (Abzhanov and Kaufman, 1999, Damen*, et al.*, 1998, Hughes and Kaufman, 2002a, Telford and Thomas, 1998). It would be of interest to know whether the second antennal segment (and its myriapod and chelicerate homologues) is allocated its identity in the canonical fashion by these hox genes or by other genes as seems to be the case in the insects. If there does appear to be a transition from the hox genes to other genes, it would be interesting to know when this occurred and indeed whether it played any role in the transition from the second antennal segment to the intercalary segment.

6.3.4 Development and evolution of intercalary segment

Functional interactions

The different genes identified in this thesis present several possible lines of enquiry into the developmental changes underlying the evolution of the insect intercalary segment

from the crustacean second antennal segment. However, to understand fully this transition, it is necessary to understand the various functional interactions involving these and other genes, and the roles they play in the patterning and differentiation of the segment. If, for example, the three genes expressed in the intercalary segment mesoderm were involved with hemocyte development, it would be necessary to know what genes they regulate, whether any of these genes are involved in the differentiation to hemocytes or even whether they are involved in regulating each other. Ultimately, a gene regulatory network of the interactions involved in the patterning and differentiation of the intercalary segment could be constructed, to describe how the segment develops.

It is important to point out that the approach I have taken in this thesis to identify candidate genes – looking for conserved gene expression associated with the intercalary segment – could miss a number of possibly important functional interactions that could have played a part in the evolution and development of the segment, namely inhibitory interactions. As was discussed above, *twi*, which is involved in the differentiation of mesoderm into muscle, does not appear to be expressed in the intercalary segment. Therefore, an important step in the evolution of the segment would have been the loss of this muscle fate. It is possible that *twi* is being repressed in the intercalary segment (although it is also possible that its expression is not being promoted). If this is the case, then any interaction inhibiting its expression is of considerable importance for understanding intercalary segment development and evolution. The same could also be true for genes in the appendage formation pathway, such as *Distal-less* (*Dll*) which could be involved in the loss of appendages on the segment. In these cases, the novel feature associated with the segment would be a lack of expression, not the presence of localised expression.

The comparative approach

The studies into intercalary segment development I have presented in this thesis outline one very important theme in evo-devo, namely the need for a comparative approach when trying to infer the developmental process underlying a conserved structure. The comparisons between *Drosophila* and *Tribolium* reaffirm the variability in early

development between the two insects, both in terms of in embryology and developmental genetics. It is important to remember that no one organism can be used as an exemplar typifying the development of a bodyplan feature, especially an organism as derived as *Drosophila*.

The studies presented here were of comparisons between *Drosophila* and *Tribolium*. For example, I showed that *kn* has a conserved pattern of expression between the two insects, suggesting a conserved function in regulating intercalary segment polarity gene expression. I proposed that *kn* could be involved in the reduction of the intercalary segment polarity gene expression based on functional interactions seen in *Drosophila*, and based on the conserved pattern of gene expression seen in *Tribolium* such a function could be conserved in the beetle. However, even if any such function were shown to be conserved between *Drosophila* and *Tribolium*, this would only be conservation between two holometabolous insects. Before being able to argue a role for *kn* in the reduction of segment polarity gene expression in the *insect* intercalary segment, it would be necessary to demonstrate conservation across the hemimetabolous and apterygote insects.

If any functional interaction involved in the development of the *Drosophila* and *Tribolium* intercalary segments are shown to not be conserved in more basal insects it would be important to know what regulatory interactions are occurring instead. For example, if *kn* is not involved in regulating intercalary segment polarity gene expression outside of *Drosophila*, it would be necessary to find out what genes do regulate the segment polarity genes in other insects, and what those genes are doing in *Drosophila*. In this way, the regulatory interactions with conserved functions in intercalary segment development across the insects could be identified, as could interactions where the *Drosophila* state is derived. This way a gene regulatory network describing the ancestral mode of development at the base of the insects (D* in figure 1.1) could be inferred. Such a model could then be used for comparisons with crustaceans, allowing the identification of the developmental changes associated with the morphological changes that occurred during the evolution of the intercalary segment.

6.4 Further work

6.4.1 Resolving pancrustacean phylogeny

The multigene phylogenetic analysis presented in chapter 3 was largely in agreement with the smaller analyses based on the nuclear datasets. Combining these smaller datasets resolved issues such as the position of the copepods. Possibly, adding more genes may help to resolve issues such as the positions of the remipedes and the cephalocarids which my analysis was unable to resolve. A common approach currently used in phylogenetics to generate large datasets is to use expressed sequence tags (ESTs). At the moment, it is unlikely that this kind of data will be generated for obscure groups like the remipedes and the cephalocarids. However, this may be feasible in the future as the costs of producing the data falls.

In the short term, a more realistic aim may be to increase the number of taxa represented in the analysis. For example, there are a number of other remipede taxa other than *Speleonectes*. Perhaps some of these will not show the artefacts in their sequence that are probably found in *Speleonectes*, and have made it so difficult to place. Also, the addition of other enigmatic taxa to the dataset, such as the mystaccocarid crustaceans mentioned earlier may help to give the more complete picture of pancrustacean phylogeny needed to understand the morphological transition which took place in the diversification of the group.

6.4.2 The development of the intercalary segment

As was demonstrated above, the studies in chapters 4 and 5 recovered a number of genes with conserved expression patterns between *Drosophila* and *Tribolium*. The expression patterns suggest that the genes have potentially very interesting roles in evolution of the intercalary segment. It is now important to investigate the functions of these genes in both *Drosophila* and *Tribolium*, to see whether they have conserved roles in the development of the segment.

The function of knot

Drosophila kn mutants lose expression of *engrailed* (*en*) and *wingless* (*wg*) and have reduced levels of *hedgehog* (*hh*) (Crozatier, *et al.*, 1999). RNAi can be used to knock *kn* out in *Tribolium* and the expression of the three segment polarity genes can be examined by *in situ* hybridisation. This would show whether the regulatory interactions seen in *Drosophila* are conserved in *Tribolium*.

However, as was discussed above, simply demonstrating a conserved role for *kn* in regulating the intercalary segment polarity genes would not be of significance for understanding the evolution of the intercalary segment. In *Drosophila*, misexpression of *ems* in the prospective mandibular segment has been argued to transform its identity to that of the intercalary segment, partly on the basis of a duplication of the smaller segment polarity gene stripes typical of the intercalary segment (Schöck, *et al.*, 2000). This provides a system to investigate whether the reduced size of the segment polarity gene stripes is dependent on *kn* expression. *In situ* hybridisation for *kn* in flies with this proposed homeotic transformation would show whether the *kn* expression domain is duplicated or expands to encompass the mandibular-maxillary segment boundary as well as the intercalary-mandibular boundary. If this is not the case, then it would be very unlikely that *kn* is involved in regulating reduced size of the segment polarity gene stripes; rather it would just be involved in establishing segment polarity gene expression in the insect intercalary segment.

The function of mesodermal genes

For the three genes showing conserved expression in the intercalary segment mesoderm, namely CG4322 (*moody*), CG32423 (*alan shepard* (*shep*)) and CG32858 (*singed* (*sn*)), it is not obvious how to assay their function. RNAi could be performed in both *Drosophila* and *Tribolium*, but there are no obvious markers to investigate phenotypes relating to the mesoderm; cuticle preps only show ectodermal features. I have suggested that these genes may be involved in the development of hemocytes. There are potential markers to investigate whether these three genes are in fact involved in the development of this major intercalary segment derivative. Nardi (2004) shows that

developing hemocytes (specifically the granular cells) in the moth *Manduca sexta* can be marked with lectins, in particular peanut agglutinin (PNA). First it would be necessary to show that this is a conserved marker for granular cells in *Drosophila* and *Tribolium*, and that it stains the cells which express *moody*, *shep* and *sn* orthologues. If this proves to be the case, then each of the three genes could be knocked out with RNAi and staining for PNA could show whether hemocyte development has been affected. A positive result would suggest a role in hemocyte development.

The function of Tribolium Blimp-1

CG5249 (*Blimp-1*) does not appear to have a role patterning the *Drosophila* head (Ng, *et al.*, 2006). However, as was discussed above, it is still necessary to investigate whether it has any such function in the *Tribolium* head. *Tc*-CG5249 could be knocked out in *Tribolium* with RNAi and larval cuticles examined for any obvious defects to the head. If, as has been argued for *Drosophila*, the gene does not have an early segmental function in the head, then there should not be any such cuticular defects. In this situation, then it seems unlikely that the gene would have a broader role in patterning the insect intercalary segment. Otherwise, it would also be useful to detail the extent of intercalary segment expression, which could be done by double *in situ* hybridisation with marker genes such as the segment polarity genes *en*, *wg* and *hh*.

Broader conservation

If any of these genes were shown to have a conserved function between *Drosophila* and *Tribolium* in patterning the intercalary segment, it would then be necessary to investigate whether or not the function is conserved more broadly in the insects. Both the beetle and fly are holometabolous insects and so a conserved function would not necessarily represent the ancestral state for the insects. The obvious starting point would be to see if the expression pattern is conserved in other insects such as *Oncopeltus*, the cricket *Gryllus bimaculatus* and the firebrat *Thermobia domestica*; three insects that span the diversity of the hemimetabolous and apterygote insects. Moreover, RNAi has been developed for *Oncopeltus* and *Gryllus* allowing assays of conserved function (Hughes and Kaufman, 2000, Miyawaki, *et al.*, 2004). If any

conservation were identified across the insects, the next obvious question would be whether the gene is present in the crustaceans, where it is expressed and what it does. For crustaceans such as *Daphnia* and *Porcellio*, *in situ* hybridisation is established (Abzhanov and Kaufman, 1999, Papillon and Telford, 2007) so it would be straight forwards to examine expression patterns.

6.5 Concluding remarks

The evolution of the insect intercalary segment from the crustacean second antennal segment provides us with a very elegant system for studying the developmental changes underlying the evolution of a novel morphological feature. Not only is there a clear segmental homology which has allowed the specific morphological transition to be defined, but there is also a diversity of insects and crustaceans amenable to developmental study making it feasible to investigate this transition. However, despite the considerable potential in studying the evolution of the intercalary segment, a great deal of uncertainty has surrounded many important issues: the phylogenetic position of the insects has been unclear and little was known about how the segment develops.

The phylogenetic and developmental studies I have presented in this thesis have begun to resolve some of these areas of uncertainty. I have helped to establish an emerging phylogenetic framework in which to view developmental changes underlying the evolution of the intercalary segment, and I have recovered a number of genes with possible roles in the evolution of several of the important derived features of the segment. I have also demonstrated that when studying such questions about bodyplan evolution, it is important to take a comparative approach. There is good reason to be optimistic that advances can be made in our understanding of this evolutionary transition.

The insect intercalary segment has the potential to be an important case study for evo-devo. The insights it can give us into how developmental change underlies morphological evolution should help to us understand the diversification of Darwin's "endless forms".

References

Abascal, F., Posada, D., Knight, R. D. and Zardoya, R. (2006). Parallel evolution of the genetic code in arthropod mitochondrial genomes. PLoS Biol. 4:711-8.

Abzhanov, A. and Kaufman, T. C. (1999). Homeotic genes and the arthropod head: Expression patterns of the *labial*, *proboscipedia*, and *Deformed* genes in crustaceans and insects. Proc. Natl. Acad. Sci. U.S.A. 96:10224-9.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic Local Alignment Search Tool. J. Mol. Biol. 215:403-10.

Amundson, R. (2005). The changing role of the embryo in evolutionary theory: Cambridge University Press.

Anderson, D. T. (1973). Embryology and Phylogeny in Annelids and Arthropods: Pergamaon Press.

Angelini, D. R., Liu, P. Z., Hughes, C. L. and Kaufman, T. C. (2005). Hox gene function and interaction in the milkweed bug *Oncopeltus fasciatus* (Hemiptera). Dev. Biol. 287:440-55.

Arthur, W. (2004). Biased embryos and evolution. Cambridge: Cambridge University Press.

Beiko, R. G., Keith, J. M., Harlow, T. J. and Ragan, M. A. (2006). Searching for convergence in phylogenetic Markov chain Monte Carlo. Syst. Biol. 55:553-65.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2007). Genbank. Nucleic Acids Res. 35:D21-D5.

Bininda-Emonds, O. R. P., Jones, K. E., Price, S. A., Grenyer, R., Cardillo, M., Habib, M., Purvis, A. and Gittleman, J. L. (2003). Supertrees are a necessary not-so-evil: A comment on Gatesy et al. Syst. Biol. 52:724-9.

Bjorum, S. M. (2006). Two genes affecting *Drosophila* gravitaxis. 47th Annual Drosophila Research Conference.

Bofkin, L. and Goldman, N. (2007). Variation in evolutionary processes at different codon positions. Mol. Biol. Evol. 24:513-21.

Boore, J. L., Lavrov, D. V. and Brown, W. M. (1998). Gene translocation links insects and crustaceans. Nature 392:667-8.

Bourlat, S. J., Juliusdottir, T., Lowe, C. J., Freeman, R., Aronowicz, J., Kirschner, M., Lander, E. S., Thorndyke, M., Nakano, H., Kohn, A. B., Heyland, A., Moroz, L. L., Copley, R. R. and Telford, M. J. (2006). Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature 444:85-8.

Boyan, G. S., Williams, J. L. D., Posser, S. and Braunig, P. (2002). Morphological and molecular data argue for the labrum being non-apical, articulated, and the appendage of the intercalary segment in the locust. Arthropod Struct. Dev. 31:65-76.

Brandley, M. C., Schmitz, A. and Reeder, T. W. (2005). Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. Syst. Biol. 54:373-90.

Briggs, D. E. G. and Bartels, C. (2001). New arthropods from the Lower Devonian Hunsrück Slate (Lower Emsian, Rhenish Massif, western Germany). Palaeontology 44:275-303.

Brown, J. M. and Lemmon, A. R. (2007). The importance of data partitioning and the utility of bayes factors in Bayesian phylogenetics. Syst. Biol. 56:643-55.

Browne, W. E., Price, A. L., Gerberding, M. and Patel, N. H. (2005). Stages of embryonic development in the amphipod crustacean, *Parhyale hawaiensis*. Genesis 42:124-49.

Brusca, R. C. and Brusca, G. J. (2003). Invertebrates. Massachusetts: Sinauer.

Budd, G. E. (2001). Tardigrades as 'stem-group arthropods': The evidence from the Cambrian fauna. Zool. Anz. 240:265-79.

Cant, K., Knowles, B. a., Mooseker, M. S. and Cooley, L. (1994). *Drosophila* singed, a fascin homolog, is required for actin bundle formation during oogenesis and bristle extension. J. Cell Biol. 125:369-80.

Carapelli, A., Lio, P., Nardi, F., van der Wath, E. and Frati, F. (2007). Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea. Bmc Evolutionary Biology 7:-.

Castoe, T. A., Doan, T. M. and Parkinson, C. L. (2004). Data partitions and complex models in Bayesian analysis: The phylogeny of Gymnophthalmid lizards. Syst. Biol. 53:448-69.

Chipman, A. D., Arthur, W. and Akam, M. (2004). A double segment periodicity underlies segment generation in centipede development. Curr. Biol. 14:1250-5.

Cohen, S. M. and Jürgens, G. (1990). Mediation of *Drosophila* head development by gap-like segmentation genes. Nature 346:482-4.

Cook, C. E., Smith, M. L., Telford, M. J., Bastianello, A. and Akam, M. (2001). *Hox* genes and the phylogeny of the arthropods. Curr. Biol. 11:759-63.

Cook, C. E., Yue, Q. and Akam, M. (2005). Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic. Proc. Natl. Acad. Sci. U.S.A. 272:1295-304.

Copf, T., Rabet, N., Celniker, S. E. and Averof, M. (2003). Posterior patterning genes and the identification of a unique body region in the brine shrimp *Artemia franciscana*. Development 130:5915-27.

Crozatier, M., Valle, D., Dubois, L., Ibnsouda, S. and Vincent, A. (1996). *collier*, a novel regulator of *Drosophila* head development, is expressed in a single mitotic domain. Curr. Biol. 6:707-18.

Crozatier, M., Valle, D., Dubois, L., Ibnsouda, S. and Vincent, A. (1999). Head versus trunk patterning in the *Drosophila* embryo; *collier* requirement for formation of the intercalary segment. Development 126:4385-94.

Crozatier, M. and Vincent, A. (1999). Requirement for the Drosophila COE transcription factor Collier in formation of an embryonic muscle: transcriptional response to notch signalling. Development 126:1495-504.

Curole, J. P. and Kocher, T. D. (1999). Mitogenomics: digging deeper with complete mitochondrial genomes. Trends Ecol. Evol. 14:394-8.

Damen, W. G. M., Hausdorf, M., Seyfarth, E. A. and Tautz, D. (1998). A conserved mode of head segmentation in arthropods revealed by the expression pattern of Hox genes in a spider. Proc. Natl. Acad. Sci. U.S.A. 95:10665-70.

Daneman, R. and Barres, B. A. (2005). The blood-brain barrier - Lessons from moody flies. Cell 123:9-12.

Darwin, C. R. (1859). On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life. London: John Murray.

de Velasco, B., Mandal, L., Mkrtchyan, M. and Hartenstein, V. (2006). Subdivision and developmental fate of the head mesoderm in *Drosophila* melanogaster. Dev. Genes Evol. 216:39-51.

Diederich, R. J., Merrill, V. K. L., Pultz, M. a. and Kaufman, T. C. (1989). Isolation, structure, and expression of *labial*, a homeotic gene of the Antennapedia Complex involved in *Drosophila* head development. Genes Dev. 3:399-414.

Diederich, R. J., Pattatucci, a. M. and Kaufman, T. C. (1991). Developmental and evolutionary implications of *labial*, *Deformed* and *engrailed* expression in the *Drosophila* head. Development 113:273-&.

Dohle, W. (1998). Myriapod-insect relationships as opposed to an insect-crustacean sister group relationship. In Arthropod Relationships, (ed. R. A. Fortey and R. H. Thomas), pp. 305-15. London: Chapman & Hall.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sorensen, M. V., Haddock, S. H. D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q. and Giribet, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745-U5.

Edgecombe, G. D. (2004). Morphological data, extant Myriapoda, and the myriapod stem-group. Contrib. Zool. 73:207-52.

Fanenbruck, M., Harzsch, S. and Wagele, J. W. (2004). The brain of the Remipedia (Crustacea) and an alternative hypothesis on their phylogenetic relationships. Proc. Natl. Acad. Sci. U.S.A. 101:3868-73.

Farzana, L. and Brown, S. J. (2008). Hedgehog signaling pathway function conserved in *Tribolium* segmentation. Dev Genes Evol 218:181-92.

Franch-Marro, X., Martin, N., Averof, M. and Casanova, J. (2006). Association of tracheal placodes with leg primordia in *Drosophila* and implications for the origin of insect tracheal systems. Development 133:785-90.

Friedrich, M. and Tautz, D. (1995). Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. Nature 376:165-7.

Gallitano-Mendel, A. and Finkelstein, R. (1998). Ectopic *orthodenticle* expression alters segment polarity gene expression but not head segment identity in the *Drosophila* embryo. Dev. Biol. 199:125-37.

Gatesy, J., Matthee, C., DeSalle, R. and Hayashi, C. (2002). Resolution of a supertree/supermatrix paradox. Syst. Biol. 51:652-64.

Giribet, G., Edgecombe, G. D. and Wheeler, W. C. (2001). Arthropod phylogeny based on eight molecular loci and morphology. Nature 413:157-61.

Giribet, G., Richter, S., Edgecombe, G. D. and Wheeler, W. C. (2005). The position of crustaceans within Arthropoda - Evidence from nine molecular loci and morphology. Crustac. Issues 16:307-52.

Gompel, N., Prud'homme, B., Wittkopp, P. J., Kassner, V. A. and Carroll, S. B. (2005). Chance caught on the wing: *cis*-regulatory evolution and the origin of pigment patterns in *Drosophila*. Nature 433:481-7.

Gould, S. J. (2002). The structure of evolutionary theory. Cambridge, Mass.: Belknap Press of Harvard University Press.

Grimaldi, D. and Engel, M. S. (2004). Evolution of the insects: Cambridge University Press.

Haas, M. S., Brown, S. J. and Beeman, R. W. (2001). Pondering the procephalon: the segmental origin of the labrum. Dev. Genes Evol. 211:89-95.

Hall, B. K. (1996). *Baupläne*, phylotypic stages, and constraint - Why there are so few types of animals. Evol. Biol. 29:215-61.

Hall, B. K. (2003). *Evo-Devo*: evolutionary developmental mechanisms. Int. J. Dev. Biol. 47:491-5.

Handel, K., Basal, A., Fan, X. and Roth, S. (2005). *Tribolium castaneum twist*: gastrulation and mesoderm formation in a short-germ beetle. Dev. Genes Evol. 215:13-31.

Handel, K., Grunfelder, C. G., Roth, S. and Sander, K. (2000). *Tribolium* embryogenesis: a SEM study of cell shapes and movements from blastoderm to serosal closure. Dev. Genes Evol. 210:167-79.

Harzsch, S. (2002). The phylogenetic significance of crustacean optic neuropils and chiasmata: A re-examination. J. Comp. Neurol. 453:10-21.

Harzsch, S. (2004). The tritocerebrum of Euarthropoda: a "non-*drosophilo*centric" perspective. Evol. Dev. 6:303-9.

Harzsch, S., Muller, C. H. and Wolf, H. (2005). From variable to constant cell numbers: cellular characteristics of the arthropod nervous system argue against a sister-group relationship of Chelicerata and "Myriapoda" but favour the Mandibulata concept. Dev. Genes Evol. 215:53-68.

Hassanin, A. (2006). Phylogeny of Arthropoda inferred from mitochondrial sequences: Strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. Mol. Phylogenet. Evol. 38:100-16.

Hassanin, A., Leger, N. and Deutsch, J. (2005). Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of Metazoa, and consequences for phylogenetic inferences. Syst. Biol. 54:277-98.

Hejnol, a. and Martindale, M. Q. (2008). Acoel development supports a simple planula-like urbilaterian. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 363:1493-501.

Horne, D. J. (2005). Homology and homoeomorphy in ostracod limbs. Hydrobiologia 538:55-80.

Huelsenbeck, J. P., Larget, B., Miller, R. E. and Ronquist, F. (2002). Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51:673-88.

Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-5.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310-4.

Hughes, C. L. and Kaufman, T. C. (2000). RNAi analysis of *Deformed*, *proboscipedia* and *Sex combs reduced* in the milkweed bug *Oncopeltus fasciatus*: novel roles for Hox genes in the Hemipteran head. Development 127:3683-94.

Hughes, C. L. and Kaufman, T. C. (2002a). Exploring the myriapod body plan: expression patterns of the ten Hox genes in a centipede. Development 129:1225-38.

Hughes, C. L. and Kaufman, T. C. (2002b). Hox genes and the evolution of the arthropod body plan. Evol. Dev. 4:459-99.

Hwang, U. W., Friedrich, M., Tautz, D., Park, C. J. and Kim, W. (2001). Mitochondrial protein phylogeny joins myriapods with chelicerates. Nature 413:154-7.

Häcker, U., Kaufmann, E., Hartmann, C., Jürgens, G., Knöchel, W. and Jäckle, H. (1995). The *Drosophila fork head* domain protein *crocodile* is required for the establishment of head structures. EMBO J. 14:5306-17.

Ikeda, Y. and Machida, R. (1998). Embryogenesis of the Dipluran *Lepidocampa weberi* Oudemans (Hexapoda, Diplura, Campodeidae): External morphology. J. Morphol. 237:101-15.

Jenner, R. A. (2006). Unburdening evo-devo: ancestral attractions, model organisms, and basal baloney. Dev. Genes Evol. 216:385-94.

Jürgens, G., Lehmann, R., Schardin, M. and Nussleinvolhard, C. (1986). Segmental organization of the head in the embryo of *Drosophila melanogaster* - a blastoderm fate map of the cuticle structures of the larval head. Rouxs Arch. Dev. Biol. 195:359-77.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. J. Am. Stat. Assoc. 90:773-95.

Klingler, M. (2004). *Tribolium* - Quick guide. Curr. Biol. 14:R639-R40.

Kraus, O. (1998). Phylogenetic relationships between higher taxa of tracheate arthropods. In Arthropod Relationships, (ed. R. A. Fortey and R. H. Thomas), pp. 296-303. London: Chapman & Hall.

Kristensen, N. P. (1981). Phylogeny of insect orders. Annu. Rev. Entomol. 26:135-57.

Lai, Z. C., Fortini, M. E. and Rubin, G. M. (1991). The embryonic expression patterns of *zfh-1* and *zfh-2*, two *Drosophila* genes encoding novel zinc-finger homeodomain proteins. Mech. Dev. 34:123-34.

Lavrov, D. V., Brown, W. M. and Boore, J. L. (2004). Phylogenetic position of the Pentastomida and (pan)crustacean relationships. Proc. R. Soc. Lond., B, Biol. Sci. 271:537-44.

Lemmon, A. R. and Moriarty, E. C. (2004). The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53:265-77.

Levine, M. (2002). How insects lose their limbs. Nature 415:848-9.

Lewis, D. L., DeCamillis, M. and Bennett, R. L. (2000). Distinct roles of the homeotic genes Ubx and abd-A in beetle embryonic abdominal appendage development. Proc. Natl. Acad. Sci. U.S.A. 97:4504-9.

Luan, Y. X., Mallatt, J. M., Xie, R. D., Yang, Y. M. and Yin, W. Y. (2005). The phylogenetic positions of three basal-hexapod groups (Protura, Diplura, and Collembola) based on ribosomal RNA gene sequences. Mol. Biol. Evol. 22:1579-92.

Lubkin, S. H. and Engel, M. S. (2005). *Permocoleus*, new genus, the first Permian beetle (Coleoptera) from North America. Ann. Entomol. Soc. Am. 98:73-6.

Lynch, J. A., Brent, A. E., Leaf, D. S., Pultz, M. A. and Desplan, C. (2006). Localized maternal *orthodenticle* patterns anterior and posterior in the long germ wasp *Nasonia*. Nature 439:728-32.

Mahaffey, J. W., Diederich, R. J. and Kaufman, T. C. (1989). Novel patterns of homeotic protein accumulation in the head of the *Drosophila* embryo. Development 105:167-74.

Mallatt, J. and Giribet, G. (2006). Further use of nearly complete, 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. Mol. Phylogenet. Evol. 40:772-94.

Mallatt, J. M., Garey, J. R. and Shultz, J. W. (2004). Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. Mol. Phylogenet. Evol. 31:178-91.

Martin, J. W. and Davis, G. E. (2001). An updated classification of the recent Crustacea.

Matsuda, R. (1965). Morphology and evolution of the insect head. Memoirs of the American Entomological Institute 4:1-334.

McGuire, J. A., Witt, C. C., Altshuler, D. L. and Remsen, J. V. (2007). Phylogenetic systematics and biogeography of hummingbirds: Bayesian and maximum likelihood analyses of partitioned data and selection of an appropriate partitioning strategy. Syst. Biol. 56:837-56.

Merrill, V. K. L., Diederich, R. J., Turner, F. R. and Kaufman, T. C. (1989). A genetic and developmental analysis of mutations in *labial*, a gene necessary for proper head formation in *Drosophila melanogaster*. Dev. Biol. 135:376-91.

Miyawaki, K., Mito, T., Sarashina, I., Zhang, H. J., Shinmyo, Y., Ohuchi, H. and Noji, S. (2004). Involvement of Wingless/Armadillo signaling in the posterior sequential segmentation in the cricket, *Gryllus bimaculatus* (Orthoptera), as revealed by RNAi analysis. Mech. Dev. 121:119-30.

Mohler, J., Mahaffey, J. W., Deutsch, E. and Vani, K. (1995). Control of *Drosophila* head segment identity by the bZIP homeotic gene *cnc*. Development 121:237-47.

Mohler, J., Vani, K., Leung, S. and Epstein, A. (1991). Segmentally restricted, cephalic expression of a leucine zipper gene during *Drosophila* embryogenesis. Mech. Dev. 34:3-9.

Nardi, F., Spinsanti, G., Boore, J. L., Carapelli, A., Dallai, R. and Frati, F. (2003). Hexapod origins: monophyletic or paraphyletic? Science 299:1887-9.

Nardi, J. B. (2004). Embryonic origins of the two main classes of hemocytes - granular cells and plasmatocytes - in *Manduca sexta*. Dev. Genes Evol. 214:19-28.

Negrisolo, E., Minelli, A. and Valle, G. (2004). The mitochondrial genome of the house centipede *Scutigera* and the monophyly versus paraphyly of myriapods. Mol. Biol. Evol. 21:770-80.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. J. R. Stat. Soc. B Met. 56:3-48.

Ng, T., Yu, F. W. and Roy, S. (2006). A homologue of the vertebrate SET domain and zinc finger protein Blimp-1 regulates terminal differentiation of the tracheal system in the *Drosophila* embryo. Dev. Genes Evol. 216:243-52.

Nie, W., Stronach, B., Panganiban, G., Shippy, T., Brown, S. and Denell, R. (2001). Molecular characterization of *Tclabial* and the 3' end of the *Tribolium* homeotic complex. Dev. Genes Evol. 211:244-51.

Nylander, J. A. A., Ronquist, F., Huelsenbeck, J. P. and Nieves-Aldrey, J. L. (2004). Bayesian phylogenetic analysis of combined data. Syst. Biol. 53:47-67.

Nylander, J. A. A., Wilgenbusch, J. C., Warren, D. L. and Swofford, D. L. (2008). AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24:581-3.

Papillon, D. and Telford, M. J. (2007). Evolution of *Hox3* and *ftz* in arthropods: insights from the crustacean *Daphnia pulex*. Dev. Genes Evol. 217:315-22.

Pavlopoulos, A. and Averof, M. (2005). Establishing genetic transformation for comparative developmental studies in the crustacean *Parhyale hawaiensis*. Proc. Natl. Acad. Sci. U.S.A. 102:7888-93.

Peterson, M. D., Rogers, B. T., Popadic, A. and Kaufman, T. C. (1999). The embryonic expression pattern of *labial*, posterior homeotic complex genes and the *teashirt* homologue in an apterygote insect. Dev. Genes Evol. 209:77-90.

Philippe, H., Snell, E. A., Bapteste, E., Lopez, P., Holland, P. W. H. and Casane, D. (2004). Phylogenomics of eukaryotes: Impact of missing data on large alignments. Mol. Biol. Evol. 21:1740-52.

Pisani, D., Poling, L. L., Lyons-Weiler, M. and Hedges, S. B. (2004). The colonization of land by animals: molecular phylogeny and divergence times among arthropods. BMC Biol. 2:1.

Popadic, a., Panganiban, G., Rusch, D., Shear, W. A. and Kaufman, T. C. (1998). Molecular evidence for the gnathobasic derivation of arthropod mandibles and for the appendicular origin of the labrum and other structures. Dev. Genes Evol. 208:142-50.

Popadic, a., Rusch, D., Peterson, M., Rogers, B. T. and Kaufman, T. C. (1996). Origin of the arthropod mandible. Nature 380:395-.

Posada, D. (2003). Selecting models of evolution. In The phylogenetic handbook - A practical approach to DNA and protein phylogeny, (ed. M. Salemi and A.-M. Vandamme), pp. 256-82: Cambridge University Press.

Posada, D. and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53:793-808.

Prpic, N. M. and Tautz, D. (2003). The expression of the proximodistal axis patterning genes *Distal-less* and *dachshund* in the appendages of *Glomeris marginata* (Myriapoda: Diplopoda) suggests a special role of these genes in patterning the head appendages. Dev. Biol. 260:97-112.

Rambaut, A. and Drummond, A. J. (2007). Tracer, version 1.4. Available from: http://tree.bio.ed.ac.uk/software/tracer/.

Regier, J. C. and Shultz, J. W. (1997). Molecular phylogeny of the major arthropod groups indicates polyphyly of crustaceans and a new hypothesis for the origin of hexapods. Mol. Biol. Evol. 14:902-13.

Regier, J. C. and Shultz, J. W. (2001). Elongation factor-2: a useful gene for arthropod phylogenetics. Mol. Phylogenet. Evol. 20:136-48.

Regier, J. C., Shultz, J. W. and Kambic, R. E. (2005). Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. Proc. R. Soc. Lond., B, Biol. Sci. 272:395-401.

Richards, O. W. and Davies, R. G. (1977). Imms' general textbook of entomology. London: Chapman and Hall.

Richards, S., Gibbs, R. A., Weinstock, G. M., Brown, S. J., Denell, R., Beeman, R. W., Bucher, G., Friedrich, M., Grimmelikhuijzen, C. J. P., Klingler, M., Lorenzen, M., Roth, S., Schroder, R., Tautz, D., Zdobnov, E. M. and Consortium, T. G. S. (2008). The genome of the model beetle and pest *Tribolium* castaneum. Nature 452:949-55.

Rogers, B. T. and Kaufman, T. C. (1996). Structure of the insect head as revealed by the EN protein pattern in developing embryos. Development 122:3419-32.

Rogers, B. T. and Kaufman, T. C. (1997). Structure of the insect head in ontogeny and phylogeny: A view from *Drosophila*. Int. Rev. Cytol. 174:1-84.

Rogers, B. T., Peterson, M. D. and Kaufman, T. C. (2002). The development and evolution of insect mouthparts as revealed by the expression patterns of gnathocephalic genes. Evol. Dev. 4:96-110.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-4.

Roonwal, M. L. (1937). Studies on the Embryology of the African Migratory Locust, Locusta migratoria migratorioides Reiche and Frm. (Orthoptera, Acrididae). II Organogeny. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 227:175-244.

Rota-Stabelli, O. and Telford, M. J. (2008). A multi criterion approach for the selection of optimal outgroups in phylogeny: Recovering some support for Mandibulata over Myriochelata using mitogenomics. Mol. Phylogenet. Evol. 48:103-11.

Sambrook, J. and Russell, D. W. (2001). Molecular cloning: a laboratory manual. Cold Spring Harbour, New York: Cold Spring Harbour Laboratory Press.

Schinko, J. B., Kreuzer, N., Offen, N., Posnien, N., Wimmer, E. A. and Bucher, G. (2008). Divergent functions of *orthodenticle*, *empty spiracles* and *buttonhead* in early head patterning of the beetle *Tribolium castaneum* (Coleoptera). Dev. Biol.

Schmidt-Ott, U. and Technau, G. M. (1992). Expression of *en* and *wg* in the embryonic head and brain of *Drosophila* indicates a refolded band of seven segment remnants. Development 116:111-&.

Scholtz, G. (1995). Head segmentation in Crustacea - an immunocytochemical study. Zoology 98:104-14.

Scholtz, G. (1998). Cleavage, germ band formation and head segmentation: the ground pattern of the Euarthropoda. In Arthropod Relationships, (ed. R. A. Fortey and R. H. Thomas), pp. 317-32. London: Chapman & Hall.

Schröder, R. (2003). The genes *orthodenticle* and *hunchback* substitute for *bicoid* in the beetle *Tribolium*. Nature 422:621-5.

Schöck, F., Reischl, J., Wimmer, E., Taubert, H., Purnell, B. A. and Jäckle, H. (2000). Phenotypic suppression of *empty spiracles* is prevented by *buttonhead*. Nature 405:351-4.

Schöniger, M. and von Haeseler, A. (1994). A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phylogenet. Evol. 3:240-41.

Seecoomar, M., Agarwal, S., Vani, K., Yang, G. and Mohler, J. (2000). *knot* is required for the hypopharyngeal lobe and its derivatives in the *Drosophila* embryo. Mech. Dev. 91:209-15.

Shippy, T. D., Guo, J. H., Brown, S. J., Beeman, R. W. and Denell, R. E. (2000). Analysis of *maxillopedia* expression pattern and larval cuticular phenotype in wild-type and mutant *Tribolium*. Genetics 155:721-31.

Shultz, J. W. and Regier, J. C. (2000). Phylogenetic analysis of arthropods using two nuclear protein-encoding genes supports a crustacean + hexapod clade. Proc. R. Soc. Lond., B, Biol. Sci. 267:1011-9.

Sinakevitch, I., Douglass, J. K., Scholtz, G., Loesel, R. and Strausfeld, N. J. (2003). Conserved and convergent organization in the optic lobes of insects and isopods, with reference to other crustacean taxa. J. Comp. Neurol. 467:150-72.

Singh, S. (1981). The Myth of Intercalary Segment in Insect Head. J. Morphol. 168:17-42.

Spears, T. and Abele, L. G. (1998). Crustacean phylogenyy inferred from 18S rDNA. In Arthropod Relationships, (ed. R. A. Fortey and R. H. Thomas), pp. 169-87. London: Chapman & Hall.

Stollewerk, A., Schoppmeier, M. and Damen, W. G. M. (2003). Involvement of *Notch* and *Delta* genes in spider segmentation. Nature 423:863-5.

Strimmer, K. and vonHaeseler, A. (1997). Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl. Acad. Sci. U.S.A. 94:6815-9.

Suchard, M. A., Weiss, R. E. and Sinsheimer, J. S. (2005). Models for estimating Bayes factors with applications to phylogeny and tests of monophyly. Biometrics 61:665-73.

Swofford, D. L. (2002). PAUP*: Phylogenetic analysis using parsimony*, version 4.0b10. Sunderland, MA: Sinauer.

Tamarelle, M. (1984). Transient rudiments of second antennae on the "intercalary" segment of embryos of Anu*rida maritima* Guer. (Collembola: Arthropleona) and

Hyphantria cunea Drury (Lepidoptera: Arctiidae). Int. J. Insect Morphol. Embryol. 13:331-6.

Telford, M. J. and Budd, G. E. (2003). The place of phylogeny and cladistics in *Evo-Devo* research. Int. J. Dev. Biol. 47:479-90.

Telford, M. J. and Thomas, R. H. (1998). Expression of homeobox genes shows chelicerate arthropods retain their deutocerebral segment. Proc. Natl. Acad. Sci. U.S.A. 95:10671-5.

Telford, M. J., Wise, M. J. and Gowri-Shankar, V. (2005). Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: Examples from the bilateria. Mol. Biol. Evol. 22:1129-36.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lweis, S. E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S. E. and Rubin, G. M. (2002). Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. Genome Biol. 3:research0088.1-.14.

Turbeville, J. M., Pfeifer, D. M., Field, K. G. and Raff, R. A. (1991). The phylogenetic status of arthropods, as inferred from 18S rRNA sequences. Mol. Biol. Evol. 8:669-86.

Turner, F. R. and Mahowald, A. P. (1979). Scanning electron microscopy of *Drosophila melanogaster* embryogenesis. III. Formation of the head and caudal segments. Dev. Biol. 68:96-109.

Uemiya, H. and Ando, H. (1987). Embryogenesis of a Springtail, *Tomocerus ishibashii* (Collembola, Tomoceridae): External Morphology. J. Morphol. 191:37-48.

Ullmann, S. L. (1964). The Origin and Structure of the Mesoderm and the Formation of the Coelomic Sacs in Tenebrio Molitor L. [Insecta, Coleoptera]. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 248:245-77.

Vachon, G., Cohen, B., Pfeifle, C., McGuffin, M. E., Botas, J. and Cohen, S. M. (1992). Homeotic Genes of the Bithorax Complex Repress Limb Development in the Abdomen of the Drosophila Embryo through the Target Gene *Distal-Less*. Cell 71:437-50.

Valentine, J. W. and Hamilton, H. (1998). Body plans, phyla and arthropods. In Arthropod Relationships, (ed. R. A. Fortey and R. H. Thomas), pp. 1-9. London: Chapman & Hall.

Veraksa, A., McGinnis, N., Li, X. L., Mohler, J. and McGinnis, W. (2000). Cap 'n' collar B cooperates with a small Maf subunit to specify pharyngeal development and suppress Deformed homeotic function in the *Drosophila* head. Development 127:4023-37.

Wiens, J. J. and Moen, D. S. (2008). Missing data and the accuracy of Bayesian phylogenetics. Journal of Systematics and Evolution 46:307-14.

Wills, M. A. (1998). A phylogeny of recent and fossil Crustacea derived from morphological characters. In Arthropod Relationships, (ed. R. A. Fortey and R. H. Thomas), pp. 189-209. London: Chapman & Hall.

Wimmer, E. A., Cohen, S. M., Jackle, H. and Desplan, C. (1997). *buttonhead* does not contribute to a combinatorial code proposed for *Drosophila* head development. Development 124:1509-17.

Wohlfrom, H., Schinko, J. B., Klingler, M. and Bucher, G. (2006). Maintenance of segment and appendage primordia by the *Tribolium* gene *knödel*. Mech. Dev. 123:430-9.

Wolff, C. and Scholtz, G. (2006). Cell lineage analysis of the mandibular segment of the amphipod *Orchestia cavimana* reveals that the crustacean paragnaths are sternal outgrowths and not limbs. Front. Zool. 4:19-32.

Yamamoto, Y., Stock, D. W. and Jeffery, W. R. (2004). Hedgehog signalling controls eye degeneration in blind cavefish. Nature 431:844-7.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367-72.

Younossi-Hartenstein, A., Tepass, U. and Hartenstein, V. (1993). Embryonic origin of the imaginal discs of the head of *Drosophila melanogaster*. Rouxs Arch. Dev. Biol. 203:60-73.

Appendix 1:
Accession numbers

Below are presented tables of accession numbers for the various sequences used in the different studies. Table A1.1 provides the accession numbers for all the sequences used in the construction of the multigene dataset for analysis of pancrustacean phylogeny. Table A1.2 provides the accession numbers for the aligned and annotated 18S and 28S ribosomal RNA sequences (downloaded from the European Ribosomal RNA database) used as templates for producing alignments of the two genes. Table A1.3 provides that accession numbers for the *Drosophila* sequences used as queries in the BLAST search of BeetleBase.

Table A1.1. Accession numbers for sequences used in assembling the multigene dataset. Accession numbers are given for all taxa from which sequences were used to construct the chimeric concatenated sequences. Accession umbers of newly sequenced data are in bold.

Taxonomic unit	Species	Nuclear genes						Mitochondrial genomes
		28S	18S	EF-1a	PolII	EF-2	H3	
Acrididae	*Melanoplus* sp.	AY125286	AF423803					
	Locusta migratoria		AF370793	AY077627			AF370817	NC 001712
Archaeognatha	*Machilis* sp.	AY521735						
	Machiloides sp.		AY084061					
	Machiloides banksi			AF137390	AF138990, AF138991, AF138992	AF240822		
	Allomachilis froggarti						AF110864	
	Nesomachilis australica							NC 006895
	Pedetontus saltator			U90056	U90041, AY305610	AY305520		
	Petrobiinae gen. sp.						AF110865	
Argulus	*Argulus* sp.	AY210804		AY305461	AY305544, AY305545, AY305546	AY305491		
	Argulus nobilis		M27187					
	Argulus americanus							NC 005935
Artemia	*Artemia* sp.	AY210805		X03349				
	Artemia salina		X01723		U10331	AF240815		
	Artemia franciscana							NC 001620

(Table A1.1 continued)

Taxonomic unit	Species	Nuclear genes						Mitochondrial genomes
		28S	18S	EF-1a	PolII	EF-2	H3	
Balanidae	*Balanus balanus*	AY520594	AY520628	AF063404			AF370818	
	Balanus crenatus	**EU914254**						
	Semibalanus balanoides		AY520626		AF138971, AF138972, AY305549	AF240817	AY520694	
	Megabalanus spinosus		AY520633					
	Megabalanus volcano							NC 006293
Blattaria	*Gromphadorhina laevigata*	AY210819						
	Gromphadorhina portentosa		Z97592					
	Periplaneta americana		AF370792	U90054	AY305602, AY305603, U90040	AY305517	AF370816	
	Periplaneta fuliginosa							NC 006076
Calanoida	*Calanus simullimus*	**EU914255**						
	Calanus pacificus		L81939					
	Eurytemora affinis			AF063408	AF138977, AY305557, AY305558	AY305497		
Campodeoidea	Campodeidae sp.	AY338649						
	Campodea tillyardi		AF173234				AF110860	
	Eumesocampa frigilis			AF137388	AF138978, AF138979, AF138980	AF240818		

(Table A1.1 continued)

Taxonomic unit	Species	Nuclear genes						Mitochondrial genomes
		28S	18S	EF-1a	PolII	EF-2	H3	
Cyclopidae	Cyclopidae sp.	AY210813						
	Eucyclops serrulatus		L81940					
	Acanthocyclops viridis		AY626999					
	Acanthocyclops vernalis			AY305458	AY305534, AY305535, AY305536			
	Mesocyclops edax			AY305470	AY305589, AY305590	AY305511		
Cypididae	Cypididae sp.	AY210815	AY210816					
	Cypridopsis japonica		AB086321					
	Cypridopsis vidua			AF063414	AF138997, AF138998, AF138999	AF240825		
Cypridinidae	*Skogsbergia lerneri*	AF363319, AF363331, AF363347	AF363297	AY305477	AY305616, AY305617, AY305618	AY305522		
	Vargula hilgendorfii	AF363317, AF363332, AF363357	AB076654					NC 005306
Daphniidae	*Daphnia occidentalis*	AF346510						
	Daphnia pulex		AF014011					NC 000844
	Scapholeberis mucronata			AF526282				

(Table A1.1 continued)

Taxonomic unit	Species	Nuclear genes						Mitochondrial genomes
		28S	18S	EF-1a	PolII	EF-2	H3	
Drosophila	*Drosophila melanogaster*	M21017	M21017	NM 165850, NM 206593	M27431	X15805	X81207	NC 001709
Entomobryomorpha	*Folsomia candida*	**EU914252**	AY555515				AY555561	
	Orchesella villosa		AY555514					
	Orchesella imitari			AY305473	AY305599, AY305600, AY305601	AY305515, AY305516		
Hexagenia	*Hexagenia* sp.	AY125276	AY121136				AY125223	
	Hexagenia limbata			AY305469	AY305584, AY305585, AY305586, AY305587, AY305588	AY305510		
Hutchinsoniella	*Hutchinsoniella macracantha*	AF370811	L81935	AF063411	AF138984, AF138985, AF138986	AF240820	AF110867	NC 005937
Japygoidea	*Parajapyx isabellae*	AY596395	AY145135					
	Heterojapyx sp.		AY555524				AY555567	
	Metajapyx subterraneus			AF137389	AF138987, AF138988	AY305503, AY305504		
	Japyx solifugus							NC 007214

(Table A1.1 continued)

Taxonomic unit	Species	Nuclear genes						Mitochondrial genomes
		28S	18S	EF-1a	PolII	EF-2	H3	
Lepas	*Lepas* sp.	**EU914256**						
	Lepas anatifera		L26516					
	Lepas anserifera			AY305466	AY305569, AY305570, AY305571	AY305505		
Lepeophtheirus	*Lepeophtheirus salmonis*		AF208263					
Lepismatidae	*Ctenolepisma longicaudata*	AY210810	AY210811					
	Ctenolepisma lineata			AF063405	AF138973, AY305553, AY305554	AY305494		
	Thermobia domestica		AF370790					NC 006080
	Lepisma saccarina		X89484					
	Lepisma sp.						AY555568	
Leptostraca	*Paranebalia longipes*	AY744899					AY744905	
	Paranebalia belizensis		AY743952					
	Nebalia sp.		L81945				AF110869	
	Nebalia hessleri			AF063413	AF138996, AY305594, AY305595	AY305513		
Limnadiidae	*Limnadopsis birchii*	AY744897						
	Limnadia lenticularis		L81934	AF063412	AF138989, AY305575, AY305576	AY305507		

(Table A1.1 continued)

Taxonomic unit	Species	Nuclear genes						Mitochondrial genomes
		28S	18S	EF-1a	PolII	EF-2	H3	
Limnadiidae (cont.)	*Limnadopsis birchii*		AY744889				AY744903	
Limulus	*Limulus polyphemus*	AF212167	L81949	U90051	U90037	AF240821	AF370813	NC 003057
Lithobius	*Lithobius* sp.	AY210825						
	Lithobius variegatus		AF000773					
	Lithobius forficatus			AF240799	AY310212, AY310213, AY310214	AY310267		NC 002629
	Lithobius sydneyensis						AF110853	
Mastigoproctus	*Mastigoproctus giganteus*	AF062989	AF005446	U90052	U90038	AF240823		AY731174
Mygalomorphae	*Aphonopelma hentzi*	AY210803						
	Aphonopelma sp.		X13457					
	Aphonopelma chalcodes			U90045	U90035			
	Atrax sp.		AF370784				AF110877	
	Ornithoctonus huwena							NC 005925
Oniscidea	*Porcellio scaber*	**EU914253**	AJ287062					
	Armadillidium vulgare		AJ287061	U90046	AY305548	AF240816		
Pauropodidae	Pauropodidae gen. sp.	AF005466	AF005451					
	Pauropodinae gen. sp.						AF110857	
	Allopauropus proximus			AY305460	AY305541, AY305542, AY305543	AY305490		

(Table A1.1 continued)

Taxonomic unit	Species	Nuclear genes						Mitochondrial genomes
		28S	18S	EF-1a	PolII	EF-2	H3	
Phalangida	*Nipponopsalis abei*	AF124975	AF124948	AF137391	AF138993, AF138994, AF138995	AF240824		
	Equitius doriae	U91503	U37003					
	Equitius sp.						AF110875	
Podura	*Podura aquatica*	AY210838	AF005452	AY305474	AY305604, AY305605	AY305518		NC 006075
Pollicipes	*Pollicipes pollicipes*	AY520616						
	Pollicipes polymerus		AY520651				AY520719	NC 005936
Pyncnogonida	*Callipallene* sp.	AY210807	AF005439					
	Colossendeis sp.	AY210809	AF005440	AF063406	AF138974, AY305555, AY305556	AY305495		
	Endeis laevis		AF005441	AF063409	AF138981, AF240882, AF240883	AF240819		
	Endeis spinosa							AY731173
	Ammothella biunguiculata						AF110874	
	Tanystylum orbiculare			AF063417	AF139013, AF139014	AF240831		

(Table A1.1 continued)

Taxonomic unit	Species	Nuclear genes						Mitochondrial genomes
		28S	18S	EF-1a	PolII	EF-2	H3	
Reptantia	*Panulirus argus*	AY210833, AY210834, AY210835	AY743955					
	Panulirus japonicus							NC 004251
	Callinectes sapidus	AF249298						NC 006281
	Libinia emarginata		AY743953	U90050	AY305572, AY305573, AY305574	AY305506		
	Homarus americanus		AF235971				AF370819	
	Eriocheir sinensis							NC 006992
	Portunus trituberculatus							NC 005037
Sacculinidae	*Sacculina carcini*	AY520622	AY265366				AY520724	
	Loxothylacus texanus		L26517	AY305467	AY305577, AY305578, AY305579, AY305580	AY305508		
Scorpiones	*Pandinus imperator*	AY210830	AY210831					
	Lychas marmoreus						AF110876	
	Centruroides sculpturatus			AF240840	AF240988, AF240989, AF240990			
	Centruroides limpidus							NC 006896
	Mesobuthus gibbosus							NC 006515

(Table A1.1 continued)

Taxonomic unit	Species	Nuclear genes						Mitochondrial genomes
		28S	18S	EF-1a	PolII	EF-2	H3	
Scutigerellidae	*Hanseniella* sp.	AY210821, AY210822	AY210823	U90049	AY305565, AF138982		AF110856	
	Scutigerella sp.	AY084064	AY336742	AF137392		AF240827		
Speleonectes	*Speleonectes gironensis*	AF370810						
	Speleonectes tulumensis		L81936	AF063416	AF139008, AF139009	AF240829		NC 005938
Spirostreptida	*Orthoporus* sp.	AY210828	AY210829					
	Orthoporus ornata			AF240802	AF240934, AF240935	AY310273, AY310274		
	Trachyiulus nordquisti			AY305479	AY305623, AY305624, AY305625, AY305626	AY305525		
	Thyropygus sp.							NC 003344
Stomatopoda	*Squilla empusa*	AY210842	L81946					
	Squilla mantis							NC 006081
	Kempina mikado		AF370802				AF110873	
	Gonodactylus sp.		L81947					
	Neogonodactylus oerstedii			AY305471	AY305591, AY305592, AY305593	AY305512		
	Harpiosquilla harpax							NC 006916

(Table A1.1 continued)

Taxonomic unit	Species	Nuclear genes						Mitochondrial genomes
		28S	18S	EF-1a	PolII	EF-2	H3	
Tigriopus	*Tigriopus californicus*	AY599492, AY599492, AF363324, AF363340, AF363350	AY599492				X52393	
	Tigriopus japonicus							NC 003979
Triops	*Triops* sp.	AY210844						
	Triops longicaudatus		AF144219	U90058	U90043, AY305622	AY305524		
	Triops australiensis						AF110870	
	Triops cancriformis							NC 004465

Table A1.2. Taxon names and accession numbers of sequences from European Ribosomal RNA database used as templates for aligning 18S and 28S rRNAs.

	Species	Accession number
18S	*Argulus nobilis*	M27187
	Artemia salina	X01723
	Callipallene gen. sp.	AF005439
	Cormocephalus monteithi	AF173249
	Daphnia pulex	AF014011
	Drosophila melanogaster	M21017
	Gromphadorhina portentosa	Z97592
	Limulus polyphemus	L81949
	Lithobius variegatus	AF000773
	Milnesium tardigradum	U49909
	Podura aquatica	AF005452
	Priapulus caudatus	X87984
	Squilla empusa	L81946
	Triops longicaudatus	AF144219
28S	*Aedes albopictus*	L22060
	Anopheles albimanus	L78065
	Caenorhabditis elegans	X03680
	Chironomus tentans	X99212
	Drosophila melanogaster	M21017

Table A1.3. Accession numbers for sequences of *Drosophila* genes used as queries in the BLAST search of BeetleBase. Where more than one isoform of a gene was used as a query sequence, accession numbers are given for all genes used in the BLAST search.

Gene	Accession number	Gene	Accession number
cnc	AAC72879	CG9520	NP 723427
croc	P32027	CG10072	NP 476980
kn	P56721	CG10130	NP 652037
wg	NP 523502	CG10521	NP 511155
hh	NP 001034065	CG10746	NP 542444
CG1322	P28166	CG10960	NP 648605
CG1444	NP 572420	CG11051	Q9VU58
CG1942	NP 610318	CG11100	NP 730768
CG3097	NP 572259	CG11188	NP 609066
CG3184	NP 572341	CG11208	NP 611460
CG3424	NP 648327	CG11415	NP 525037
CG3597	NP 608674	CG11546	NP 652028
CG3732	NP 611692	CG11798	NP 611013
CG3762	NP 652004	CG12177	NP 572911
CG3879	NP 523724	CG12708	NP 727875
CG4261	NP 732097	CG13037	NP 524104
CG4280	Q27367	CG13475	NP 652614
CG4322	NP 569970	CG13651	NP 651343
CG4501	NP 524698	CG13894	NP 612054
CG5059	NP 649239	CG15162	NP 523597
CG5249	NP 647982	CG15211	NP 572653
CG5575	NP 523833	CG17786	NP 651231
CG5663	NP 650192	CG17932	NP 652627
CG5840	NP 650632	CG18375	NP 788423
CG5893	NP 524066	CG31150	NP 732076
CG6096	NP 524511	CG31607	NP 723350
CG6117	P16912	CG31629	NP 001097107
CG6207	NP 648448	CG31811	Q9NGC3
CG7271	NP 649041	CG32372	NP 729265
CG8036	NP 649812	CG32423	NP 729054
CG9005	NP 610688	CG32434	NP 996129
CG9148	NP 477392		NP 730594
CG9171	NP 723117	CG32858	NP 511076
CG9238	NP 648708	CG33099	NP 788714
CG9415	NP 524722		
	NP 726032		

Appendix 2:
Primer sequences

Below are presented tables of the primer sequences used in the different studies. Table A2.1 gives the sequences of the primers used to amplify 28S rRNA. Table A2.2 gives the sequences of the primers used to amplify partial cDNAs of *Tribolium* genes. Tables A2.3 give the primer sequences used for sequencing reactions. Table A2.4 give the primer sequences used for amplifying probe synthesis templates.

Table A2.1. Sequences of primers used to amplify fragments of 28S rRNA.

5’ – Forward	U178	GCACCCGCTGAAYTTAAGCA
	U212	GGAAAAGAAACTAACMRGGA
	U427	TCGGGTTGTTTGRGARTGCA
	U541	AGAGAGAGTTCAARAGKRCGTGA
	U940	GGCCACCCTCTCGACCGT
	U1148	GACCCGAAAGATGGTGAACTA
	U1372	ACGATCTCAACCTATTCTCAAACT
	U1640	CCTGAAAATGGATGGCGCT
	U1846	AGGCCGAAGTGGAGAAGGGTT
	U2229	TACCCATATCCGCAGCAGGTCT
	U2562	AAACGGCGGGAGTAACTATGA
	U2771	AGAGGTGTAGGATARGTGGGA
	U3119	TTAAGCAAGAGGTGTCAGAAAAGT
	U3139	AAGTTACCACAGGGATAACTGGCT
3’ – Reverse	L538	ACGTACTTTTGAACTCTCTCTTCA
	L1149	CATACTTCACCATCTTTCGGGT
	L1344	CAAGGCCTCTAATCATTCGCT
	L1642	CCAGCGCCATCCATTTTCA
	L1964	AATATTAACCCGATTCCCTTTCG
	L2230	AGACCTGCTGCGGATATGGGT
	L2450	GCTTTGTTTTAATTAGACAGTCGGA
	L2630	GGGAATCTCGTTAATCCATTCA
	L2984	CTGAGCTCGCCTTAGGACACCT
	L3358	AACCTGCGGTTCCTCTCGTACT
	L3449	GATTCTGACTTAGAGGCGTTCA

Table A2.2. Sequences of primers used to amplify *Tribolium* partial cDNAs. The approximate size of the amplified fragment is given, as this is the size of the probe synthesised from the fragmnent.

Tc-cnc	Forward	GAT TAC AGC TAT ACG AGT CGG	
	Reverse	GTC AGC CAG ACT CAA AAT CTG	750 bp
Tc-croc	Forward	ATG CAT ACG ATT TTC ACC GAA	
	Reverse	CTC CTT CTC GCG GAG GGC GTC	500 bp
Tc-kn	Forward	GGA ATA CAG TAT AGG CTG CAG	
	Reverse	ATG CCT GGG AAT GAG CTT TTG	900 bp
Tc-lab	Forward	ACA TAC CCA TCG GAT AAC TAC	
	Reverse	CCT TTT GAC TTG CAT CCA CTT	550 bp
Tc-wg	Forward	GGA TGC AGG GAA ACT GCC TTC	
	Reverse	AAC GCA AGT ATG TAT GGT TCT	1000 bp
Tc-hh	Forward	TAT AAC CAG GAC ATC GTC TTC	
	Reverse	ACT GTC AAT GGT CGC GTA ACA	800 bp
Tc-CG1322	Forward	GTC CGA GTC CGT TCG TTA ATT	
	Reverse	CAC GTG GTG CTT GTG CTT GAA	900 bp
Tc-CG1444	Forward	ACT GAT GGA ATT GGC AAA GCC	
	Reverse	GGA GTA TTC GGA GTT CAA GTC	500 bp
Tc-CG3184	Forward	GAG GCG TGG ATT TGT GCC TTT	
	Reverse	TGA CCA ATC AGC CCC ATA AGC	500 bp
Tc-CG3732	Forward	AAT TTC GCC CGC CGT AAC AAC	
	Reverse	GTC AGA CTC GTG CTC CTT ATA	500 bp
Tc-CG4261	Forward	GCC CTT CGT GAA ATA ATC ACC	
	Reverse	TAT AAA AGG ACA TGC GGC ATG	1000 bp
Tc-CG4280	Forward	CCG ATT CCG ATG TAC ATC GAG	
	Reverse	CAC TGG ATA CCA TAA TTC CCC	650 bp
	Forward		
Tc-CG4322	Reverse	ATG TTC TGC TTC ATC GTC CTC	
		GTA GAT TAG GAT GTA GCC CAG	800 bp
	Forward		
Tc-CG4501	Reverse	GCG AAT GCT CTT AAA GGC TCG	
		AAG TTC CTT CAA ACG CCC CGT	700 bp
	Forward		
Tc-CG5249	Reverse	TAC CCC CTG AAG AAG AAG GAC	
		ACA ACT CGT CGT CTT CCA ATG	450 bp
	Forward		
Tc-CG5575	Reverse	GAC AAT TAC GTT GTG ACT CCG	
		CGG TCC TGA CAA ATG CCT GAT	800 bp
	Forward		
Tc-CG5840	Reverse	TCT CGA GTG GTT CGA GTG ATG	
		ATT CGT GTC CCT CAC CAT TTG	300 bp

(Table A2.2 continued)

Tc-CG5893	Forward	ATG AAC GCC TTC ATG GTC TGG	
	Reverse	ATA CAT AAC TGG GAC CGG CCT	650 bp
Tc-CG6207	Forward	CTG TAT GTA ATA ACC CCG ACC	
	Reverse	CTT CTT CGT TTG CGT GTG CCA	650 bp
Tc-CG9148	Forward	AAG GAC GGA TAC ATT TCG CGG	
	Reverse	CTT TGT TAG CTG TTC GTC GGC	650 bp
Tc-CG9238	Forward	GTT CGA GTC ATG ACG GAA CCT	
	Reverse	TTC ATT CTC GTT GCA ACG GAA	450 bp
Tc-CG9520	Forward	GGC AAG AGG TGC AAC AAG TTG	
	Reverse	GAG ACA CTT CCC CAT TTC GAC	400 bp
Tc-CG10072	Forward	CCA ACG TGT AGT GTC ATA GCT	
	Reverse	AGT CGC AAC TTC GGA AAC ATC	700 bp
Tc-CG10130	Forward	ACT GTT AGG CAG AGG AAG ACC	
	Reverse	AGA CCT AGT GTA CTT TCC CCA	200 bp
Tc-CG11208	Forward	ATC CAC TAC ATC GGC ATG CGT	
	Reverse	CAG CCT AGC CCC TAA TAA CAA	700 bp
Tc-CG11415	Forward	TAC ATC GGG CTG TAT GTC TTG	
	Reverse	ATG GAA GAA CGG ATT TCC GGT	400 bp
Tc-CG11546	Forward	GTG AAA AGC GAA GAC GCT CTC	
	Reverse	CTC AAT CAC CAA GTC GTC AGT	600 bp
Tc-CG11798	Forward	ATA ACC AGG AAG CTC TAC GGC	
	Reverse	CTC ATG CCG CGT AAA TAG GTC	900 bp
Tc-CG12177	Forward	TAC ACC GCA GTT GGA AAC ATC	
	Reverse	GAA TGC CTC AGC ATC AAA CTG	500 bp
Tc-CG13037	Forward	AAC TTC GGT GTG GGC CGA TTA	
	Reverse	CTC TTC CGC AAC CCT GTA TTT	400 bp
Tc-CG13475	Forward	TTC CAG GGA CTC GTC TCC AAC	
	Reverse	TTC TTG TTT CCG CTT GGC CGT	300 bp
Tc-CG18375	Forward	CTT GAA GGC GAG CTG GAA TTG	
	Reverse	CCA CCA TTC CCT TTC ATT CTC	500 bp
Tc-CG31150	Forward	GCG TCC GTT CTG TAC ATC AAG	
	Reverse	CAT GTA CGA GTG CAC GAA ACG	700 bp
Tc-CG31811	Forward	AGA ATG AAG AGT AGT GGG GTG	
	Reverse	GTT TGA ACT GCA GAC AGC ATC	700 bp

(Table A2.2 continued)

Tc-CG32372	Forward	AAG ATC TCC TTC AGC AAG CTG	
	Reverse	CTG AAG GCC CAC AAT CTC TTG	900 bp
Tc-CG32423	Forward	ATA AGA GGA CTG AAT CCG ACC	
	Reverse	ACC GTC GGC AAA CAA GAC CAA	500 bp
Tc-CG32434	Forward	TCG GGA ATG CAA GTC GAT GTT	
	Reverse	TAG CTC AGC CTC AAT CCT CAA	800 bp
Tc-CG32858	Forward	ATC CAC GTT GAT GCC AAC ATC	
	Reverse	TTC GCC CCG TTC GAC TTG AAT	800 bp

Table A2.3. Sequences of primers used for sequencing reactions. Primers are named according to the cloning vector polymerase sites to which they were designed to anneal.

SP6	GAT TTA GGT GAC ACT ATA
T7	TAA TAC GAC TCA CTA TAG GG
T3	AAT TAA CCC TCA CTA AAG GGA

Table A2.4. Sequences of primers used for amplifying probe synthesis template. Primers pBS-A and pBS-E were used with pCR II-TOPO, pGEM-T Easy, pFLC-1 and pBS vectors, and primers OTf and OTr with the pOT2 vector.

pBS-A	CTA TGA CCA TGA TTA CGC CAA G
pBS-E	TAA CGC CAG GGT TTT CCC AGT
OTf	AAT GCA GGT TAA CCT GGC TTA TCG
OTr	AAC GCG GCT ACA ATT AAT ACA TAA CC

Appendix 3:

Drosophila clone references

Below is presented a table of the specific *Drosophila* clones ordered from the BDGP for each gene.

Table A3.1. Name of *Drosophila* clones used for each gene.

Gene	Clone
cnc	LD12047
croc	RH24787
kn	RE03728
lab	RE63854
CG32423	RH63980
CG4322	RE06985
CG32858	RH62992